Faculdade de Medicina de Lisboa

PADRÕES DE COMORBILIDADE DOS DOENTES COM ALTA HOSPITALAR NA REGIÃO DE LISBOA E VALE DO TEJO ENTRE 2009 E 2011

António José Correia Botelho de Sousa

Mestrado em Epidemiologia

Lisboa 2015

Faculdade de Medicina de Lisboa

PADRÕES DE COMORBILIDADE DOS DOENTES COM ALTA HOSPITALAR NA REGIÃO DE LISBOA E VALE DO TEJO ENTRE 2009 E 2011

António José Correia Botelho de Sousa

Orientador: Prof. Doutor Paulo Jorge Nogueira

Mestrado em Epidemiologia

A impressão desta dissertação foi aprovada pelo Conselho Cientifico da Faculdade de Medicina de Lisboa em reunião de 25 de Novembro de 2014

Agradecimentos

Os meus agradecimentos a todos os colegas e amigos que de uma forma ou de outra contribuíram para a criação das condições que permitiram a elaboração desta dissertação.

O meu agradecimento especial ao meu orientador Prof. Doutor Paulo Nogueira pela paciência, a disponibilidade e o apoio que me deu, sempre incondicional e pronto, apesar dos seus inúmeros afazeres.

Um agradecimento também ao meu querido amigo Prof. Doutor Joaquim Silveira Sérgio por me ter instigado a envolver-me neste projecto. Parte do que aqui está é fruto da sua persistência e motivação. Também uma palavra para outro querido amigo o Mestre Dr. Ricardo São João pela ajuda sempre pronta e dedicada que me deu. Bem haja por isso.

Finalmente um agradecimento muito especial á minha família, a minha mulher e os meus filhos Paula, Francisco e Rita pela paciência de me aturarem ao longo destes longos anos e pelo apoio permanente e dedicado que me proporcionaram. Sem eles nada seria possível.

Finalmente, mais do que agradecer, quero dedicar este trabalho á memória do meu pai que sempre esteve do meu lado.

Resumo

O desafio da compreensão da comorbilidade continua a constituir um desafio em aberto para os serviços de saúde continua a ser uma questão sem resposta, não obstante os vários desenvolvimentos ao nível do registo na base de dados de Grupos de Diagnósticos Homogéneos (GDHs).

Uma gestão eficiente da comorbilidade das doenças (principalmente as de cariz crónico) para além de ser indissociável dos pacientes e médicos, também está ligada ao modelo de organização e á orientação dos cuidados de saúde.

O presente trabalho tem como objectivo identificar os principais padrões de comorbilidade em indivíduos da Região de Lisboa e Vale do Tejo, com ênfase nas altas hospitalares que ocorreram no ano de 2009, 2010 e 2011 em internamentos com duração inferior a um ano.

Palavras-Chave: comorbilidade, multimorbilidade, analise factorial, regressão logística

Abstract

The challenge of understanding comorbidity remains an open problem to the health services, being a question without answer, aside of the improvement of the database records of the Diagnosis Related Groups (GDH in portuguese). An efficient management of the disease comorbidity (with special focus in the chronic diseases) is linked to the organizational model, and the policy of care, but also with the patients and the doctors. The aim of this work was to identify the main comorbidity paterns of Lisbon and Tagus River Valley, focusing on Hospital discharged patients during 2009-2011, with stay under one year, and with the knowledge of these patterns study the relationships between co and multimorbidity and the patients characteristics.

Keywords: comorbidity, multimorbidity, factor analysis, logistic regression

Conteúdo

Lista de Figuras

Lista de Quadros

Capítulo 1

Introdução

1.1 O Conceito de comorbilidade

O conceito de comorbilidade é introduzido na literatura médica pelo epidemiologista clínico norte-americano Alvan Feinstein em 1970 num trabalho hoje considerado histórico [22].

Este investigador (4 Dezembro 1925/25 Outubro 2011) é autor de vinte livros publicados e de mais de quatrocentos trabalhos, que lhe granjearam renome mundial, embora também alguma controvérsia á volta da sua figura.

De entre esses trabalhos é no citado artigo que Feinstein primeiro aborda o conceito de comorbilidade, curiosamente num caso particular de doente com febre reumática.

Assim, segundo o autor, "num doente com uma doença principal definida (doença índice) o termo comorbilidade refere-se a qualquer condição adicional coexistente".

O facto de se omitir a análise e a classificação da comorbilidade conduz a vários problemas na análise estatística.

"A omissão conduz a incorrecções no cálculo das taxas de mortalidade para uma população geral, ou de taxas de letalidade para uma doença especificada. Em particular a omissão da comorbilidade pode causar comparações espúrias durante o planeamento e a avaliação do tratamento de doentes com diagnósticos aparentemente idênticos" (op. citada).

A comorbilidade pode alterar o curso da doença em pacientes com o mesmo diagnóstico, uma vez que modifica a duração da mesma, e os aspectos ligados à terapêutica, ao prognóstico e ao próprio desfecho da doença inicial.

Para além destes efeitos directos na evolução clínica, a comorbilidade desempenha igualmente um papel nas decisões, que pode alterar a própria

classificação diagnostica.

Estas decisões prendem-se com aspectos como:

- Atribuição de sintomas em doentes com poli-patologia;
- A selecção de manifestações iniciais da doença índice

Este conceito inicial foi mais alargado, existindo hoje uma outra abordagem a que certos autores denominam de mutlimorbilidade [69] cujas diferenças podemos assim sintetizar:

1. Comorbilidade indica uma condição médica, num doente, que causa, é causada, ou está de qualquer forma relacionada com qualquer outra condição no doente
2. Multimorbilidade indica uma condição médica existindo simultaneamente, mas de forma independente, com outra condição num doente;

A diferença não é despiciente, e a forma como as duas entidades são encaradas, tem levado a abundante literatura sobre a matéria, abrangendo as várias disciplinas médicas.

1.2 A Abordagem clássica

Podemos de uma forma sintética dizer que a abordagem de Feinstein suscitou o interesse em encontrar uma forma de "medir" a comorbilidade, numa perspectiva essencialmente preditiva, isto é, encontrar um indicador a partir do qual fosse possível prever a evolução do doente.

Os mais geralmente aceites são [15]:

O **Índice de Charlson** [13] eventualmente o mais conhecido prevê a mortalidade a dez anos de um doente com um conjunto de várias patologias (até 22).

A cada condição é atribuída uma pontuação de 1, 2, 3 ou 6, dependendo do risco de morte associada a cada uma. A soma das respectivas pontuações dá-nos o risco final.

Existem algumas variantes do índice de Charlson, tais como os de: **Charlson/Deyo** [18], **Charlson/Romano** [57], **Charlson/Manitoba** [57] e **Charlson/D'Hoores** [19].

As doenças são classificadas da seguinte forma:

- 1 cada: Enfarte do miocárdio, Insuficiência cardíaca congestiva, doença vascular periférica, demência, doença cerebrovascular, doença

crónica do fígado, doença do conectivo, úlcera, doença crónica do fígado;

- 2 cada: Hemiplegia, doença renal severa ou moderada, diabetes, diabetes com complicações, tumor, leucemia, linfoma;
- 3 cada: Doença herpética severa ou moderada;
- 6 cada: Tumor maligno, metástases ou SIDA

A importância deste índice advém da possibilidade do tratamento de dada doença poder estar complicada na sua relação custo eficácia pela presença de um valor elevado.

Podemos analisar um destes casos, mais especificamente no tratamento da diabetes, num dos múltiplos trabalhos do próprio Feinstein [39].

Trata-se de um estudo realizado em doentes adultos portadores de diabetes mellitus não insulinodependentes.

A partir de registos médicos, os autores efectuaram um "follow-up" por cinco (5) anos numa coorte inicial de 188 doentes tratados durante os anos de 1959-1962, e cuja diabetes foi diagnosticada dentro dos 6 meses anteriores ao "tempo zero" considerando este a data da alta do evento que motivou o internamento.

Foi desenvolvida uma classificação especial, a fim de categorizar a comorbilidade dos doentes no "tempo zero" como, forte ou ligeira e, dividir a mortalidade severa nos tipos vascular e não vascular, com a severidade ordenada como fraca, moderada ou severa.

A taxa de letalidade aos 5 anos após o "tempo zero" era 40% (76/188) para todos os doentes; mas gradientes distintos de letalidade estavam associados á idade, ao tipo de comorbilidade e, em particular, ao grau de severidade da comorbilidade.

A taxa de letalidade aos 5 anos depois do "tempo zero" era de 7% em 41 doentes com comorbilidade ligeira, 33% em 79 doentes com comorbilidade moderada, e 69% em 68 doentes com comorbilidade severa.

Dos 68 doentes com comorbilidade severa inicial, 53% faleceram mais tarde da mesma doença, ou de doença associada; em 79 doentes com comorbilidade moderada a taxa foi de 13%.

A morte foi causada por causas vasculares em 52% dos 77 doentes que apresentavam comorbilidade vascular inicialmente, em 7% dos doentes que não apresentavam comorbilidade vascular e em 2% de 41 doentes sem comorbilidade forte.

Entre os sobreviventes a 5 anos a ocorrência de novos eventos vasculares (as chamadas complicações diabéticas) estava directamente relacionada com os mesmos aspectos de idade e comorbilidade que pareciam afectar os óbitos.

Estes dados indicam que o resultado final dos doentes, com diabetes, não insulinodependente do adulto depende do tipo e da severidade funcional da comorbilidade presente quando a diabetes é detectada.

Uma análise apropriada da comorbilidade, embora anteriormente omitida das avaliações estatísticas da diabetes, é um pré-requisito para avaliar os resultados de diferentes terapêuticas.

O **Índice de Elixhauser** [20] foi desenvolvido utilizando dados administrativos a partir de uma base de dados de doentes internados em estabelecimentos da Califórnia.

Este índice desenvolveu uma medida de comorbilidade com uma lista de trinta (30) diagnósticos baseados na ICD-9CM.

O **Cumulative Rating Scale** foi desenvolvido em 1968 por B. S. Linn [45].

As comorbilidades identificadas pelo índice de Elixhauser estão fortemente associadas á mortalidade intra-hospitalar e incluem doenças agudas e crónicas.

Foi um índice em certa medida revolucionário, na medida em que proporcionou aos médicos a oportunidade de calcular o número e a severidade das doenças crónicas, dentro do quadro da comorbilidade dos doentes.

A própria denominação do índice significa "Avaliação Cumulativa separada de cada um dos sistemas biológicos".

Em linhas gerais calcula-se como:

- 0 corresponde á ausência de doença
- 1 Doença ligeira ou patologia anterior
- 2 Doença necessitando de terapêutica
- 3 Doença causadora de incapacidade
- 4 Doença aguda necessitando de terapêutica de emergência

O índice varia entre 0 e 56.

Note-se, a propósito, que trabalhos mais recentes [35] alargaram a utilidade deste índice á prática da Medicina Geral e Familiar como referimos mais adiante.

O **Índice de Kaplan-Feinstein** [39] criado em 1973 é baseado no estudo do efeito das doenças associadas a doentes sofrendo de diabetes tipo II.

Os pormenores deste índice podem ser apreciados no artigo referido.

O **Índice de Doença Coexistente (ICED)** foi inicialmente desenvolvido em 1993 por Greenfield, sendo posteriormente adaptado [37]; foi pensado tendo os doentes com cancro em vista, tendo o seu campo sido posteriormente alargado a outras áreas como, por exemplo, a Nefrologia [53].

O **Índice GIC (Geriatric Index of Comorbidity)** desenvolvido em 2002 [60].

O **Índice FCI (Funcitonal Comorbidity Index)** desenvolvido em 2005 [28].

O **Índice TIBI (Total Illness Burden Index)** desenvolvido em 2007 [46].

1.3 Evolução do conceito

A breve resenha da secção anterior levou-nos através de uma primeira, e imediata, reacção da comunidade científica, no sentido de utilizar a noção de comorbilidade para fins de previsão, sobretudo do risco de morte do doentes com certos valores à partida.

Várias especialidades médicas exploraram esta fonte, como as ligadas á Saúde Mental [62], mas igualmente as ligadas á problemática dos idosos [60], á Ortopedia [27], á Reumatologia [55], e aos Cuidados Intensivos [13], entre outras.

Actualmente existe, por vezes, alguma imprecisão no uso do termo comorbilidade.

Desta maneira certos autores referem-se á comorbilidade como, a presença de doenças num doente condicionadas por mecanismos patogénicos conhecidos a uma doença inicial, enquanto a multimorbilidade é definida como a presença em simultâneo de várias doenças num mesmo doente, não necessariamente ligadas entre si por um mecanismo patogénico identificado.

Como acima se referiu deve-se essencialmente ao trabalho de H.C. Kraemer e M. van der Aker (op. citada) a clarificação dos conceitos, da forma acima referida.

A investigação à volta dos mesmos é vasta.

Vamos referir alguns resultados da literatura, bem como dum nosso trabalho preliminar efectuado em 2011, e que não se encontra publicado.

Assim referimos:

Um estudo realizado ao longo de dez anos em doentes portadores de seis doenças crónicas demonstrou que praticamente metade dos doentes idosos com artrite também tinham hipertensão, 20% tinham doença cardíaca e 14% diabetes tipo II [11] ;

Em doentes idosos com nefropatia crónica a frequência de doença coronária é de 22% superior, e os novos eventos coronários 3.4 vezes mais frequentes do que em doentes com função renal mantida [4];

Num estudo conduzido em 483 doentes obesos verificou-se que a dimensão da obesidade, associada a outras doenças, era superior entre as mulheres do que entre os homens.

Os investigadores concluíram que 75% de doentes obesos tinham doenças concomitantes, nomeadamente dislipidemia, hipertensão e diabetes tipo II.

Verificou-se igualmente que entre os obesos jovens (18 aos 29 anos) em 22% dos homens e 43% das mulheres foram encontradas mais de duas doenças associadas [8].

A fibromialgia é uma doença com comorbilidade com outras doenças, incluindo mas não limitada a: depressão, ansiedade, cefaleias, síndrome de bexiga irritável, síndrome de fadiga crónica, lúpus eritematoso sistémico, artrite reumatóide [71] migraine e perturbações de pânico [36].

Num pequeno estudo exploratório por nós realizado com a colaboração de Ricardo São João, integrado na Cadeira de Análise Multi-factorial do curso de Especialização em Epidemiologia da Faculdade de Medicina de Lisboa em 2011 (não publicado), tivemos a oportunidade de analisar as altas hospitalares de 2009 na Região de Lisboa e Vale do Tejo, num total de 719603 registos.

Uma abordagem empírica permitiu obter os resultados do quadro 1.1

Quadro 1.1: Comorbilidade em Lisboa e Vale do Tejo 2011

Comorbilidades			
2	3	4	5
Nº de doentes			
181106	9016	2368	774

Os nosso resultados, revelaram cerca de 25% de casos com duas doenças e valores de cerca de 1,2% no segundo grupo e abaixo de 1% nos restantes grupos.

Uma vez que se tratou de uma análise empírica não podemos adiantar grande significado aos mesmos.

Contudo ficam como ponto de partida.

Deste breve resumo podemos verificar que a questão da comorbilidade se apresenta como um elemento que pode influenciar o desempenho das instituições e, nomeadamente a forma de prestação de cuidados.

1.4 Aspectos complementares da comorbilidade

Para terminar este capítulo introdutório vamos referir dois aspectos que derivam duma perspectiva mais alargada a literatura hoje questiona justamente o impacto nomeadamente nos Cuidados de Saúde Primários e na qualidade de vida dos doentes.

1.4.1 Comorbilidade e Cuidados de Saúde Primários

Num artigo de 2003, Barbara Starfield et all. [65] estudaram a importância do fenómeno de comorbilidade na utilização dos Cuidados de Saúde Primários (CSP).

Um dos desafios resultante do sucesso, quer da Medicina Preventiva quer da Medicina Curativa, é o aumento da extensão das comorbilidades, ou seja o aparecimento de doenças aparentemente não relacionadas.

No estudo referido os autores partiram de uma análise retrospectiva de dados administrativos, curiosamente não relacionados com idosos, inscritos num serviço de prestação de cuidados de saúde.

A análise feita aos registos, em termos de visitas a Médicos de Família e a Especialistas, quer por doenças índice, quer por outras condições associadas levou os autores a concluir que: "O número de visitas, quer a médicos de Cuidados Primários, quer a Especialistas, encontra-se fortemente associada com o grau de comorbilidade. No caso de doenças menos comuns, os especialistas são mais provavelmente escolhidos do que os generalistas para a condição inicial, mas não para a comorbilidade; investigação prévia já tinha demonstrado que os médicos de Cuidados Primários referenciam aparentemente mais doentes com situações pouco comuns, do que doentes com situações muito comuns" [65].

O resultado não se podendo classificar como inesperado e, estando aliás de acordo com outros estudos, como o de Kuhlthau [44] curiosamente este em crianças, não deixa de abrir portas para uma reflexão sobre a eventual adequação dos serviços de Cuidados Primários a uma demanda mais exigente.

1.4.2 Comorbilidade e Qualidade de vida

Outra vertente que interessa valorizar, no contexto desta temática, é a da qualidade de vida dos doentes portadores de níveis pesados de comorbilidade.

Referimos aqui, igualmente á guisa de exemplo, um estudo interessante de Martin Fortin et. all [25] em que os autores analisam um conjunto de 753 doentes com diferentes graus de comorbilidade no sentido de avaliarem o impacto da comorbilidade e da multimorbilidade na Qualidade de Vida dos doentes.

Para tal fizeram uma revisão sistemática de bases de dados electrónicas (Medline e Embase) no período de 1990 a 2003.

Daqui concluíram, embora com várias limitações metodológicas que os autores identificam e, que comprometem em certa medida a validade externa e mesmo a validade interna do estudo, que parece existir uma relação inversa entre o grau de comorbilidade e a Qualidade de Vida dos doentes.

1.5 Conclusão

Ao terminar esta brevíssima introdução ao problema da comorbilidade podemos afirmar que:

- Existe um problema, amplamente reconhecido, de confusão entre os conceitos de comorbilidade e multimorbilidade que convém tentar clarificar;
- Inicialmente tentou aproveitar-se o conceito no sentido de construir índices que pudessem fornecer indicações de previsão sobre o destino final dos doentes, de acordo com pontuações iniciais;
- Contudo estudos posteriores demonstraram uma dimensão mais vasta do conceito, envolvendo aspectos como o impacto nos serviços de saúde, nomeadamente em Cuidados de Saúde Primários, e mesmo na Qualidade de Vida dos doentes.

Curiosamente esta situação estende-se mesmo ao universo das crianças.

No capítulo seguinte vamos aprofundar mais um pouco os conceitos de comorbilidade e de multimorbilidade e os vários aspectos dos respectivos constructos.

Capítulo 2

Comorbilidade e Multimorbilidade

2.1 Introdução

A comorbilidade como foi definida por Feinstein [22] apresentada no capítulo anterior pressupõe a existência de uma doença índice e de uma, ou mais, associadas.

A aplicação do conceito foi igualmente apresentada anteriormente e, revelou a grande amplitude das suas potencialidades em várias disciplinas médicas.

Os autores demonstraram um interesse especial nesta matéria; uma referência especial para van den Akker [69] deve ser feita que num artigo de 1996 apresentou uma revisão da literatura sobre o assunto centrada no período de 1996 a 1994.

Nesse artigo os autores verificaram que, "a ocorrência de condições médicas é um fenómeno comum, com tendência a crescer e acarreta múltiplas consequências".

Os autores verificaram igualmente a ocorrência de outras definições, o que tornava o conceito ambíguo.

Assim, "a fim de diminuir a indiferenciação no que concerne à terminologia, propuseram a distinção entre dois termos":

comorbilidade, conforme a definição original

multimorbilidade definida como a ocorrência de múltiplas doenças, agudas ou crónicas no mesmo doente.

Para os dois conceitos propõem uma classificação em três categorias:

1. comorbilidade/multimorbilidade simples; co-ocorrência das doenças sejam elas coincidentes ou não;

2. comorbilidade/multimorbilidade associativa; sem implicação de causalidade;

3. comorbilidade/multimorbilidade; implicando uma relação causal entre doenças coexistentes.

Como podemos verificar a introdução do novo conceito de multimorbilidade alargou o espectro de situações que possam vir a ser objecto de estudo. Ainda segundo as palavras do autor [69], "estudando a multimorbilidade, podemos encontrar padrões que possam indicar determinantes".

É esse justamente o objectivo do nosso trabalho.

A prevalência da multimorbilidade foi estudada por vários autores quer no âmbito hospitalar, quer no âmbito de cuidados primários.

A lista é profundamente exaustiva, pelo que referiremos aqui os mais relevantes na nossa pesquisa [24, 64].

No caso dos doentes hospitalares [23] num estudo realizado em Santander, Espanha encontrou multimorbilidade em 18% (IC a 95% 15,8-18,1) das altas hospitalares; destas 27,4% (IC a 95% 24,2-ti0,6) cumpriam mais de duas categorias da definição.

Noutro extremo no estudo de Schneider [64] realizado em Zurique, Suíça o número de doentes, com mais de 65 anos, portadores de duas doenças, de acordo com dados administrativos ou registos médicos variaram entre os 86,5% e os 90%.

Este exemplo demonstra bem como é difícil neste momento encontrar uma resposta padronizada, em termos de medida mas até mesmo em termos conceptuais, para a implicação do conceito de multimorbilidade na prática clínica e epidemiológica.

Isso mesmo foi enfaticamente referido por Daniel Campbell-Scherer [9] num editorial publicado em 2010.

Neste texto o autor aborda os seguintes tópicos a considerar no impacto da multimorbilidade na Medicina Baseada na Evidência:

- A multimorbilidade modifica o efeito dos tratamentos se entendidos para uma doença isolada, pelo que as "guidelines" pensadas nessas condições poderão apresentar resultados inesperados no mundo real, no tratamento de doentes com multimorbilidade.

- Os ensaios clínicos realizados sem levar em consideração o fenómeno apresentarão resultados enviesados, com evidente compromisso da sua

validade interna, mas também no que concerne á sua generalizabilidade (validade externa).

Nas palavras do autor "existe uma dessincronização entre os cuidados focados no indivíduo, versus os cuidados focados na doença".

Conclui assim que "a multimorbilidade representa a próxima fronteira na evolução da Medicina Baseada na Evidência".

2.2 O modelo conceptual

2.2.1 Abordagem elementar

Vários autores [7, 67] debruçaram-se sobre um modelo conceptual que permitisse visualizar os conceitos.

Cynthia M. Boyd [7] definiu um modelo relativamente simples para os conceitos de comorbilidade, como definido por Feinstein [22] e o de multimorbilidade definido no trabalho de van den Akker [69].

Na Figura 2.1 representamos o modelo para o conceito de comorbilidade.: doença índice, com uma ou mais doenças afectando o seu curso e tratamento.

A comorbilidade tem sido estudada e encarada na prática clínica na perspectiva de uma doença índice, e uma ou mais doenças concorrentes devem ser consideradas.

Estas doenças podem afectar o curso e o tratamento da doença índice em vários graus (daí a diferença entre as várias linhas).

Este quadro pode criar planos de tratamento dissonantes para cada patologia e tornar-se pesado em doentes com várias doenças coexistentes [7].

Na figura 2.2 representamos a forma como podemos visionar o diagrama conceptual da multimorbilidade.

A perspectiva da multimorbilidade pode ser usada para tratar doentes com múltiplas condições.

Estas incluem as doenças tradicionais, mas também podem reflectir situações como deficiência, traumatismos, etc. que caem fora do tradicional modelo de doença.

Estas condições podem sobrepor-se em vários graus.

A intersecção das mesmas pode acontecer dentro de um contexto de saúde biológica, tal como em circunstâncias psicológicas de um indivíduo (exemplo afectos positivos).

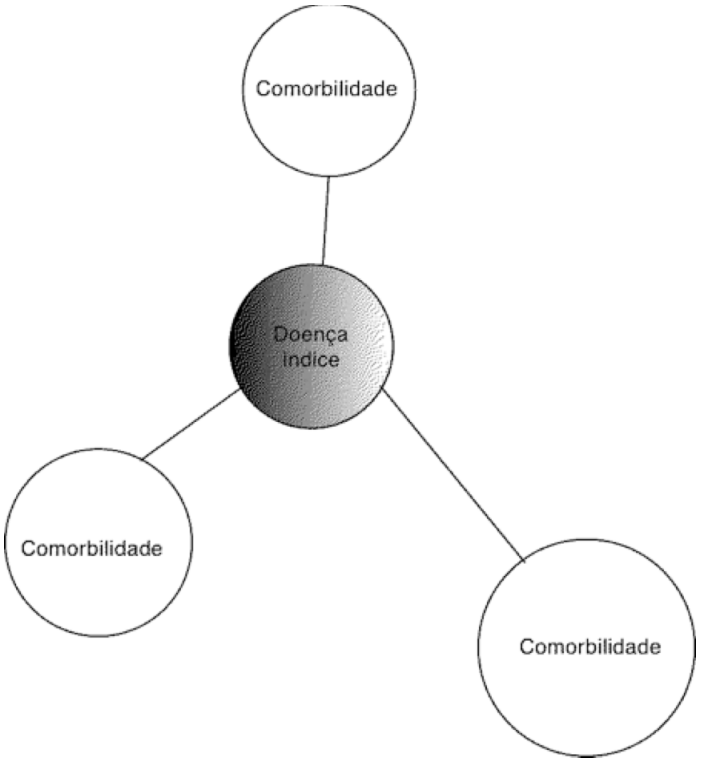


Figura 2.1: Modelo conceptual de comorbilidade

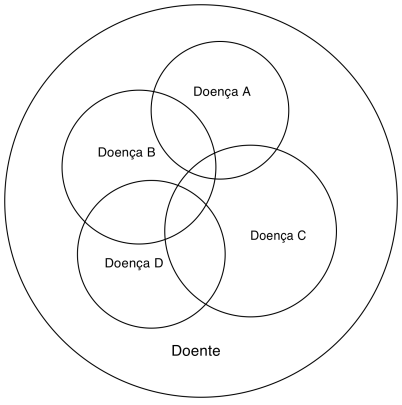


Figura 2.2: Modelo conceptual de multimorbilidade

As situações de multimorbilidade também para dados indivíduos se desenrolam dentro das suas circunstâncias económicas, social, culturais, educacionais e ambientais, e estas afectarão o desenvolvimento da situação de multimorbilidade.

O indivíduo portador de multimorbilidade também apresenta valores individuais e prioridades para a sua vida e para os seus cuidados de saúde , as quais deverão ser expressas e levadas em linha de consideração aquando da elaboração de planos de saúde [7].

2.2.2 Abordagem mais avançada

Complementarmente ao modelo anteriormente analisado Valderas [67] propõe um modelo mais avançado o qual contempla não só o que o autor denomina de "fardo da doença", mas igualmente um outro constructo que o autor identifica como "complexidade do doente".

Da interacção destes constructos resulta um doente complexo que deve ser analisado, mesmo na perspectiva terapêutica, de uma forma global.

Outrossim, o doente funcionaria como uma aglomeração de situações clínicas que poderiam ser isoladamente abordadas resultando numa visão atomizada que nada beneficiaria o doente.

Seguindo um exemplo citado pelo próprio Valderas, podemos imaginar uma doente, sexo feminino, 60 anos de idade, diabética, hipertensa, com depressão, de uma minoria étnica e com baixa literacia, que simultaneamente cuida do seu marido, incapacitado na sequência de acidente vascular cerebral.

Analisada na óptica do psiquiatra, este iria valorizar a sua depressão, considerando a diabetes e a hipertensão como comorbilidades.

Na óptica do médico de medicina geral e familiar seria considerada a multimorbilidade, dado que este avaliaria com o mesmo peso, a diabetes, a hipertensão e a depressão.

O seu fardo de doença, medido por qualquer dos índices disponíveis, seria determinado pela presença das várias doenças, tomando a sua relativa severidade em conta.

Finalmente a sua complexidade como doente seria igualmente modulada pelo seu nível cultural de base, as suas limitações linguísticas, e pela sua situação pessoal como um todo, incluindo as condições de vida, e, não menos importante a sua situação de cuidadora do marido.

A figura 2.3 , adaptada do mesmo artigo, apresenta estes conceitos de forma mais visual.

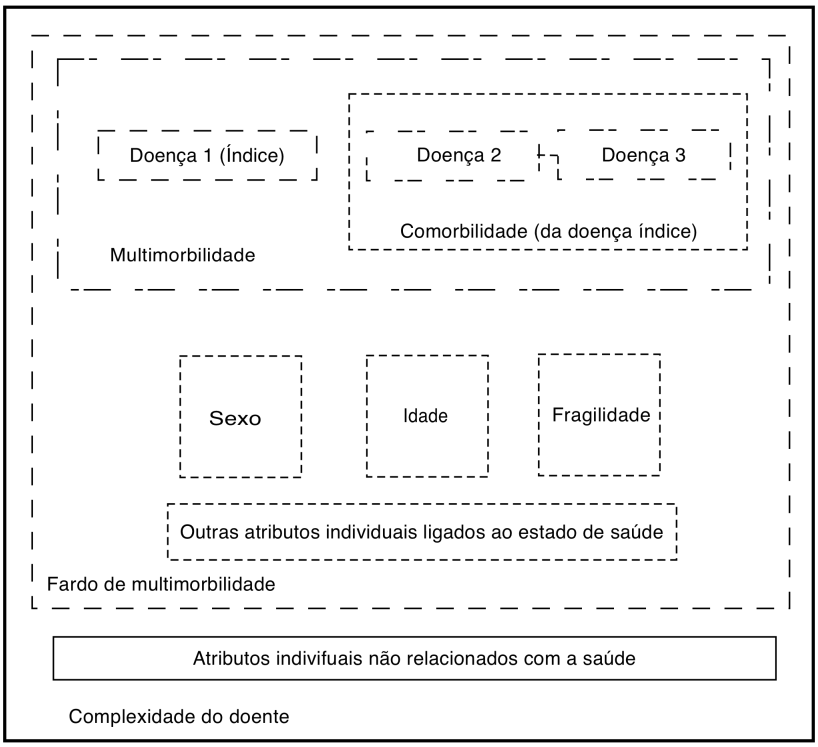


Figura 2.3: Os constructos de co e multimorbilidade

Seguindo a nossa exposição de perto a linha de pensamento do autor, dizemos então que existem três áreas de investigação onde é útil aplicar e desenvolver estes constructos, a saber:

- A Clínica
- A Epidemiologia
- O Planeamento em Saúde

A nível da clínica, o constructo a utilizar será determinado pela sua capacidade de auxiliar no tratamento do doente.

Embora a noção da complexidade do doente seja relevante, em todos os aspectos, para o seu tratamento, o constructo de comorbilidade com o seu foco numa doença principal é bastante útil na área das especialidades, focadas que estas estão na doença principal que o motiva o recurso às mesmas.

Já a nível de cuidados primários o constructo de multimorbilidade relevase mais útil, atendendo a que se atende ao doente na sua globalidade, mais do que aos aspectos particulares da sua situação patológica.

No campo da Epidemiologia o foco é colocado no estudo da associação entre exposição e evento.

Estes constructos permitem fornecer informação que possa levar a uma melhor estimação da incidência e da prevalência [30]

Sob o ponto de vista do planeamento o conhecimento de padrões de comorbilidade e de multimorbilidade poderão ajudar a uma melhor alocação dos recursos, e a uma intervenção mais eficaz e mais eficiente.

2.2.3 Relação entre comorbilidade e multimorbilidade

Continuando a seguir o pensamento de Valderas [67] vamos, por razões de simplicidade, imaginar uma estrutura com 2 doenças num mesmo indivíduo.

Duas questões se põem:

- Existe algum traço de união etiológico entre elas, e que impacto têm no cuidado a prestar ao doente?

- Existe um conjunto alargado de factores que condicionam a saúde de um indivíduo, desde a sua constituição genética a factores ambientais e até opções políticas.

É previsível que as doenças se organizem em *clusters* num doente, se elas partilharem um padrão de influencias, ou se a resiliência ou a vulnerabilidade do indivíduo for alterada.

Contudo, devemos estar prevenidos para o facto de outros factores poderem explicar a formação destes *clusters*.

Existem 3 vias principais pelas quais diferentes doenças podem ser encontradas no mesmo indivíduo: acaso, viés de selecção ou associação causal.

A comorbilidade que acontece por acaso ou por viés de selecção, embora sem ligação causal, pode ser importantes por poder alterar a percepção dessas mesmas associações causais.

Duas doenças podem ocorrer somente por acaso.

Imaginemos que uma tem uma prevalência de 4% e a outra de 5%.

Nestas condições esperaríamos que elas ocorressem em conjunto em 0,2% dos casos (0,4*0,5=0,002).

Uma associação com importância teria portanto de se afastar deste valor de uma forma estatisticamente significativa.

Os vieses de selecção são outra explicação alternativa.

O seu estudo está largamente difundido, em livros e textos diversos [31, 32, 49, 58].

A sua importância justifica aqui uma nota mais extensa

2.2.4 Vieses

Para além dos autores citados, um artigo de 2004 de Delgado-Rodrigues [17] faz uma sistematização bastante completa do conceito de viés.

Recordemos que os vieses ocorrem quando as características da amostra seleccionada difere das características da população alvo, sobre a qual se pretende fazer posteriores inferências.

Assim, na definição dos autores, o viés de selecção "é o erro introduzido quando a população em estudo não representa a população alvo".

Desta maneira, vemos que "o viés de selecção pode ser controlado quando as variáveis que influenciam a selecção podem ser medidas em todos os indivíduos em estudo e:

1. São anteriores quer á exposição, quer ao acontecimento;

2. A distribuição conjunta dessas variáveis (mais a exposição e o acontecimento) é conhecida em toda a população alvo;

3. A probabilidade de selecção para cada nível dessas variáveis é conhecida"

Pode ser introduzido em cada passo do estudo:

- Desenho
- Má definição da população elegível
- Falta de precisão do quadro amostral
- Procedimentos diagnósticos incorrectos na população alvo
- Implementação

Os autores apresentam uma longa lista dos vários vieses, pelo que nos iremos cingir aos mais frequentes remetendo a lista completa para o artigo citado.

Destes (vieses de selecção), que ocorrem por definição inapropriada da população elegível citamos:

Viés dos cuidados de saúde (Healthcare acess bias) Acontece quando seleccionamos casos a partir de registos de serviços de saúde, uma vez que a população que frequenta estes serviços não é representativa da população em geral (vão aqueles que mais cuidados têm com a saúde).

Viés do saudável Presente quando se seleccionam casos a partir de populações previamente rastreadas para efeitos de trabalho (os que estão no activo serão os mais saudáveis).

Viés de Neyman (Incidence-prevalence bias) Acontece quando são seleccionados uma série de sobreviventes, se a exposição estiver associada ao factor prognóstico.

Como exemplo podemos considerar o caso da associação entre o consumo do tabaco e o enfarte agudo do miocárdio (EAM), sendo os casos recolhidos uma semana após o ataque.

Se os doentes fumadores com EAM morrerem com mais frequência, os remanescentes apresentarão taxas mais baixas de consumo de tabaco, subestimando a associação entre consumo de tabaco e EAM.

Viés de exclusão Observa-se quando os controlos com a exposição são excluídos, enquanto os casos com as mesmas condições são mantidos no estudo.

Exemplifica-se com o estudo da associação da reserpina com o cancro da mama.

Os controlos com doença cardiovascular (uma situação comum e associada com o uso da reserpina) foram excluídos do estudo, mas não os casos, conduzindo assim a uma associação espúria entre a reserpina e o cancro da mama.

Falta de precisão do quadro amostral Embora possamos considerar vários vieses incluídos nesta categoria, tais como os vieses de citação, de disseminação, análise post-hoc, publicação, etc. o mais frequente é sem dúvida o viés de não aleatorização.

Viés de não aleatorização Sempre que não se garante a representatividade dos vários grupos na amostra.

O exemplo mais flagrante será o dos casos das entrevistas telefónicas, em que somente parte da população (a que tem telefone) é consultada. De referir ainda que na fase de implementação do estudo é ainda possível a introdução de vieses, a saber:

O **viés de perda para o follow-up (Loss to follow-up)** , o **Viés de informação em falta em análise multivariada (Missing information on multivariate analysis)**, e o **Viés de falta de resposta (Non-response bias)**.

Referimos em último lugar o **viés de Berkson**, talvez dos primeiros a ser estudado e revelado por este autor em 1946 [5].

Neste texto o autor chamava a atenção para o erro que se comete quando se pretendem fazer inferências sobre uma população, seleccionando como amostra desta doentes hospitalares.

Assim, em estudo de caso controlo, a probabilidade de hospitalização é diferente para os casos e para os controlos.

A situação é sobejamente conhecida e é aqui referida somente à guisa de curiosidade.

2.2.5 Modelos etiológicos de comorbilidade

Regressando á questão da multimorbilidade, ainda segundo o autor que estamos a seguir definem-se 4 modelos de associação etiológica entre, neste caso, as duas condições:

- Causa directa
- Factores de risco associados
- Heterogeneidade
- Independência

Todos estes modelos podem estar subjacentes a quadros de co e multimorbilidade, pelo que se justifica conhecê-los um pouco melhor.

O primeiro exemplo da figura 2.4 é da não existência de associação etiológica entre as doenças.

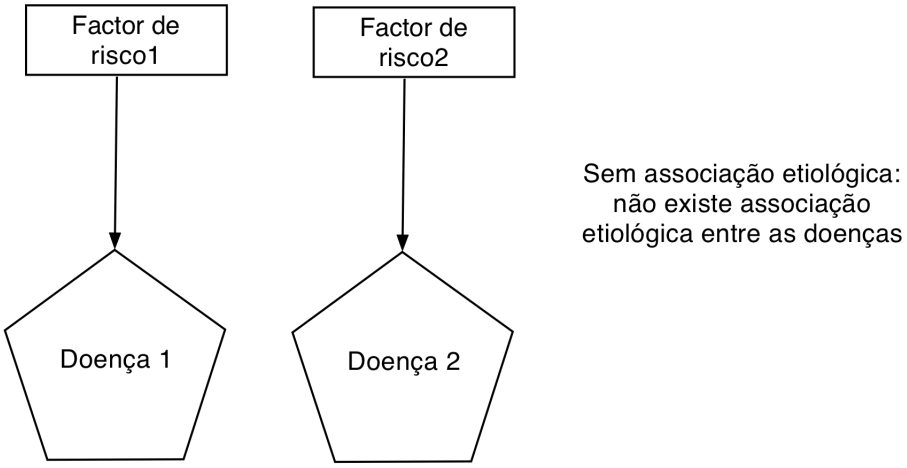


Figura 2.4: Sem associação etiológica

O segundo exemplo da figura 2.5 é o da associação causal: neste caso, uma doença é directamente responsável pela outra.

Será o caso da diabetes e da retinopatia diabética, sendo que a primeira é causa da segunda. Note-se que esta situação se pode entender igualmente no caso de um tratamento (ex: terapêutica anticoagulante para a fibrilhação auricular provoca uma hemorragia digestiva).

O modelo da figura 2.6 é o dos factores de risco associados.

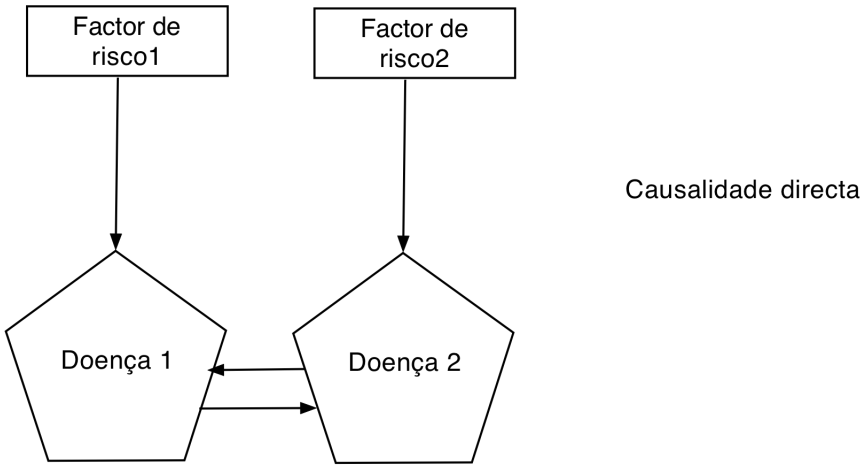


Figura 2.5: Associação directa

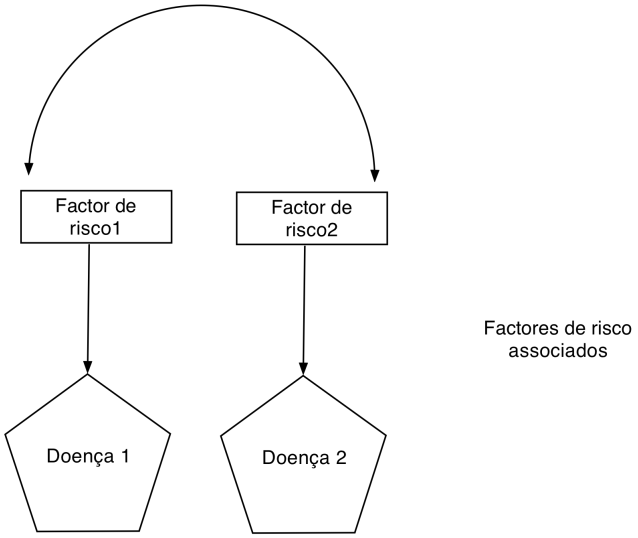


Figura 2.6: Risco associado

Neste caso o factor de risco para a doença 1 está correlacionado com o factor de risco para a outra doença, tornando a ocorrência simultânea das duas doenças mais provável.

Um bom exemplo será o consumo de tabaco (Risco1) e o consumo de álcool (Risco2).

O primeiro é um factor conhecido para a doença crónica obstrutiva pulmonar, o segundo para a doença hepática crónica, tornando portanto a ocorrência das duas doenças em simultâneo mais provável.

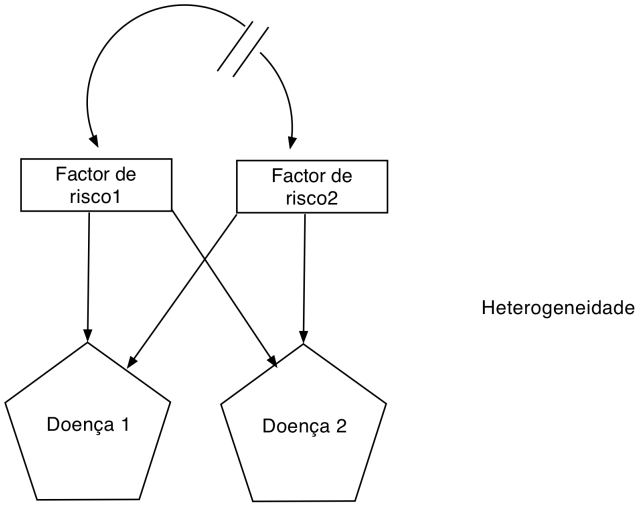


Figura 2.7: Heterogeneidade

O modelo de heterogeneidade da figura 2.7 apresenta as diferenças do modelo anterior.

Como podemos observar os factores de risco não estão correlacionados entre si, mas cada um deles é capaz de provocar doenças associadas com o outro factor de risco.

Um bom exemplo deste modelo será o caso do consumo de tabaco e da idade.

São factores independentes (não correlacionados entre si), mas como sabemos ambos estão envolvidos na génese de doenças cardiovasculares (hipertensão, enfarte, etc.) e de várias doenças neoplásicas.

Finalmente o modelo da figura 2.8 apresenta o modelo de independência; neste modelo a presença de dois quadros clínicos corresponde á acção de um terceiro.

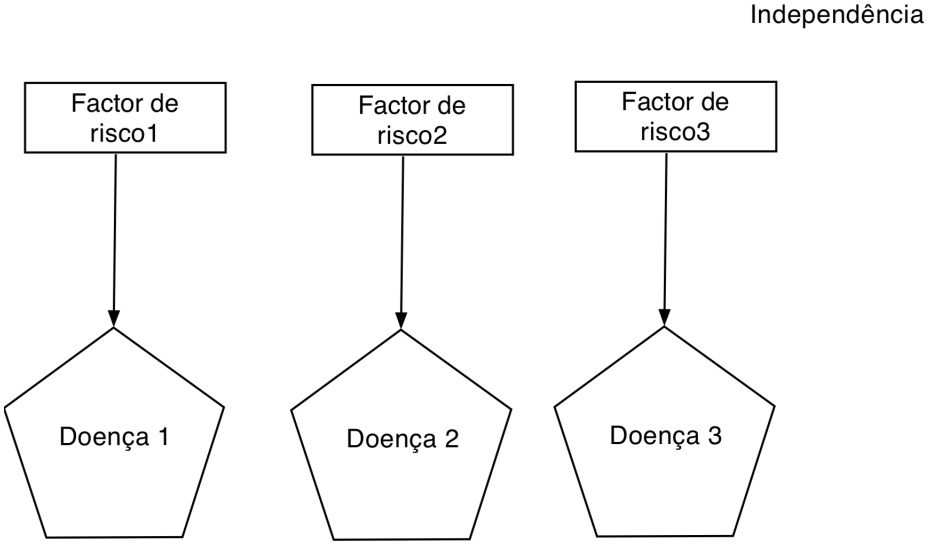


Figura 2.8: Independência

Podemos exemplificar este quadro com o exemplo da hipertensão e da cefaleia terem ambas um feocromocitoma subjacente.

Este conjunto de modelos propostos por Valderas [67] como acima se disse dão bem a imagem da complexidade das interelações entre factores de risco (exposições) e doenças (eventos).

Esta abordagem, um pouco diferente da abordagem convencional, faz especialmente sentido quando encaramos os aspectos da comorbilidade e da multimorbilidade no diagnóstico, tratamento e prognóstico posterior do doente.

Á guisa de exemplo podemos presentar o caso da diabetes, ou melhor do doente diabético: a doença isquémica cardíaca, os factores de risco cardiovasculares (hipertensão, hipercolesterolemia, etc.) e a diabetes são habitualmente acompanhadas em conjunto, em termos de cuidados primários, uma vez que elas apresentam vários aspectos em comum, no que concerne á orientação do doente.

Apresentamos um conjunto de interacções entre a diabetes e um conjunto de comorbilidades.

- Diagnóstico mais fácil devido ás doenças associadas:
  - Muitos diabéticos fazem fundoscopia permitindo assim detectar uma retinopatia não diabética

- Diagnóstico mais difícil devido a uma doença coexistente:
  - Muitos Diabéticos têm tolerância diminuída á dor podendo tornar mais difícil o diagnóstico de enfarte do miocárdio
- Tratamento indicado para doença coexistente:
  - Exercício físico indicado a um doente com DPOC (beneficia a diabetes)
- Efeito antagónico na doença coexistente:
  - Corticoesteróides prescritos para uma doença pulmonar agravam a situação da diabetes
- Prognóstico modificado pela presença de uma doença coexistente:
  - Mortalidade associada com diabetes aumenta na presença de insuficiência arterial periférica
- Prognóstico não modificado:
  - Diabetes não afectada por hipotiroidismo

Uma compreensão alargada dessas interacções é importante a fim de melhorar a prestação de cuidados.

Uma palavra ainda para o, cada vez mais premente, mundo das doenças crónicas, onde os efeitos da co e multimorbilidade têm grande impacto na saúde, e nos cuidados aos doentes.

Mais um exemplo do autor e muito sugestivo; um traumatismo, mesmo que ligeiro, da perna direita, pode levar um doente destes á imobilização, afectando desta forma o controlo da diabetes, e tornando aparente uma até aí desconhecida lesão osteoartrósica do joelho contra lateral.

Feitas estas considerações vamos no capítulo seguinte apresentar algumas estratégias, que a literatura nos propõe, par tentar encontrar padrões de comorbilidade e multimorbilidade que nos possam ser úteis para implementação de melhores diagnósticos e atitudes terapêuticas.

Capítulo 3

Medir a comorbilidade e a multimorbilidade

3.1 Introdução

Apresentados que foram os conceitos de comorbilidade e de multimorbilidade nos capítulos anteriores, vamos agora fazer uma revisão das propostas que a literatura nos faz para a sua medição.

Consultando as grandes bases de dados documentais (PubMed, Google Scholar) verificamos a existência de inúmeros artigos quer para os descritores "Measuring Comorbidity" 107.000 e "Measuring Multimorbidity" 4.030.

Refinando um pouco a pesquisa encontrámos um estudo recente (2012) de Tilahun Nigatu Haregu [30] que apresenta uma revisão da literatura sobre esta matéria, que nos ajudará a melhor compreender o estado actual do problema.

De acordo com estes autores, "não existe dúvida acerca da importância da medição da multimorbilidade e da comorbilidade. Contudo, muitas das revisões de medições de comorbilidade e de multimorbilidade estão limitadas aos índices de comorbilidade/multimorbilidade, apesar da presença de outras relevantes dimensões das mesmas. As medições epidemiológicas e os indicadores que são aplicadas em multimorbilidade e em comorbilidade não estão bem estipulados de forma sistematica".

Rever a capacidade de medir a multimorbilidade e a comorbilidade contribuirá para avançar na compreensão das mesmas.

Desta forma parece claro a possibilidade de abordar esta problemática a dois níveis; num deles a análise centra-se no estudo dos vários índices, conforme definidos em capítulos anteriores.

Noutro nível, este de mais interesse para nós, a estratégia consiste em

"identificar subpopulações com múltiplas doenças, interagindo entre si, a fim de lhes proporcionar cuidados de saúde adequados [54].

A primeira estratégia afasta-se do objectivo deste trabalho, pelo que nos iremos centrar na segunda estratégia referida.

Lembremos que, ainda de acordo com os autores citados [54] a investigação para identificar estas populações têm utilizado modelos multivariados de análise de regressão, a fim de caracterizar características individuais (como idade, sexo, doenças, etc.) que possam prever um evento de interesse (como a hospitalização).

Comparado com este método, as técnicas de *data minning* fornecem-nos uma oportunidade de identificar, de modo empírico grupos de doentes com padrões semelhantes de multimorbilidade.

Uma dessas técnicas, a Análise de Clusters, refere-se a um método de classificação que é utilizado a fim de descobrir grupos ou *clusters* de indivíduos altamente semelhantes dentro da base de dados.

A Análise de Cluster*s* é muito utilizada na psicologia, na sociologia, nos estudos de mercado, mas e uma forma mais discreta na investigação em serviços de saúde.

Pode e, tem sido utilizada, na perspectiva de encontrar "grupos" de patologias que se associam entre si, ou na de encontrar grupos de doentes que partilhem diagnósticos que necessitem de cuidados especiais.

De acordo com Schafer et all [63] "se tentarmos compreender a doença num quadro de multimorbilidade ficamos perante um labirinto de possibilidades. Todas as doenças estão mais ou menos estatisticamente associadas umas com as outras. Se começarmos em qualquer ponto deste labirinto e não tivermos um guia para encontrar o nosso caminho, facilmente nos perderemos. Por essa razão torna-se importante descobrir a estrutura subjacente na distribuição da combinação das doenças, i. e. quais os caminhos que nos possam conduzir através do labirinto da multimorbilidade".

Para além de artigos relacionados com situações especiais, como o uso de álcool e acidentes que recorrem a serviços de urgência [52], da insuficiência cardíaca nos idosos [61], ou da dor torácica [33] vamos analisar em pormenor a estratégia seguida em quatro artigos: o de Cornell [14], o de Miera [23], o de Schaffer [63] e o de Newcomer [54].

Os três primeiros referem-se á análise de modelos multimorbilidade multivariados, na perspectiva de agrupar patologias semelhantes entre si e o ultimo na procura de grupos de doentes que partilhem patologias semelhantes.

3.2 Descrição de estudos relevantes

3.2.1 Multimorbidity clusters: clustering binary data from multimorbidity clusters: clustering binary data from a large administrative medical database

Neste trabalho da autoria de Cornel et all [14], os autores começam por "descrever e ilustrar a aplicação da Análise de Clusters para identificar grupos de multimorbilidade relevantes".

Para tanto seleccionaram 45 doenças crónicas em doentes de cuidados primários (N=1327328) com duas ou mais doenças crónicas, consultados no Veterans Health Administration.

Observaram assim a presença de seis clusters: *Cluster Metabólico, Cluster de Obesidade, Cluster Hepático, Cluster Neurovascular, Cluster de Stress* e *Cluster de Diagnóstico Dual.*

A técnica utilizada foi a da construção de uma matriz de proximidade de *pxp* (as variáveis correspondentes ás patologias), as quais forma previamente convertidas em variáveis binárias (0/1) representando a presença ou ausência da patologia.

Procedeu-se então a uma análise de clusters hierárquica, baseada no algoritmo de Lance-Williams, implementado no algoritmo *agnes* do software estatístico *R*, conforme descrito pelos seus criadores Kaufman e Rousseuw [59].

Utilizou-se o coeficiente de Jaccard como medida de dissemelhança.

Metodologicamente devemos referir que os autores chegaram ás 45 patologias referidas, as quais se apresentam no anexo do trabalho citado, através de um painel de peritos de 7 serviços de saúde, o qual elaborou a lista final.

Os resultados do estudo foram os seguintes:

1645314 doentes cumpriram os critérios elegíveis na primeira visita.

Destes 94,6% eram do sexo masculino.

A média de idades era de 62,4 anos com um DP=13,6.

76% eram brancos, 17% eram Afro Americanos, 6% hispânicos e 1% pertenciam a outros grupos étnicos.

O número médio de diagnósticos por doente era de 3.49 com um DP=2.22 e variavam entre 0 e 20 doenças.

A prevalência das 45 doenças crónicas constam da Tabela 3.1

Como acima dissemos os autores encontraram 6 clusters com significado clínico constituídos da seguinte forma:

Quadro 3.1: Doentes por grupo

0	1	2
101814 (6.12%)	216118 (13.1%)	1327382 (80.7%)

Obesidade Cluster com osteoartrite, lombalgias, hipertrofia da próstata, refluxo esogástrico e obesidade;

Metabólico Cluster com diabetes, hipertensão, hiperlipidemia, e doença isquémica cardíaca;

Neurovascular Cluster com doença vascular periférica, trombose intra-arterial, AVC, D. de Alzheimer e convulsões;

Fígado Cluster com hepatite B, hepatite C, doença crónica do fígado e HIV;

Diagnóstico Dual Cluster com abuso de substâncias, dependência do álcool, esquizofrenia, doença bipolar;

Misto Ansiedade-Depressão Cluster com depressão e distúrbios de ansiedade

O cluster composto de grupos de diagnóstico com o maior grau de associação era o *Cluster Metabólico*.

Dos doentes com 2 ou mais doenças crónicas 1088774 (83%) caíram dentro de uma das 15 possíveis combinações das 4 doenças dentro do cluster (ex: hipertensão e hiperlipidemia, hipertensão e diabetes, hipertensão, hiperlipidemia e diabetes, etc.).

A prevalência para os doentes com as quatro doenças crónicas neste cluster foi de 0.05033.

A emergência deste cluster não constituiu uma surpresa, uma vez que este já tinha sido anteriormente referido como "O Quarteto Mortal" [16].

O segundo cluster mais prevalente era o *Cluster de Obesidade*; 711272 (54%) de todos os doentes que tinham duas ou mais doenças caiam numa das 31 possíveis combinações das 5 doenças dentro deste cluster (osteoartrite, lombalgias, hipertrofia da próstata, refluxo esogástrico e obesidade).

A prevalência para doentes com as 5 doenças neste cluster era de 0.00038.

O reconhecimento deste cluster é importante considerando a prevalência destas doenças e o potencial para efeitos adversos no tratamento de doenças

neste cluster. Por exemplo o uso de anti-inflamatórios não esteróides para o tratamento da osteoartrite pode agravar as queixas de refluxo esogástrico.

As doenças nos dois clusters com menos prevalência, o *Neurovascular* (186881; 14%) e o *Fígado* (48466; 4%) têm fortes laços epidemiológicos em termos de factores de risco ou de transmissão de doença.

A prevalência para os doentes com todas as doenças que definem estes clusters era de 0.00025 e 0.00004 para o *Neurovascular* e para o *Fígado* respectivamente.

A análise da frequência mostrou que 280711 (21%) dos doentes com 2 ou mais doenças crónicas caiam num dos subgrupos de doenças do cluster *Misto Ansiedade-Depressão* e 153962 (12%) caiam num dos subgrupos de doenças do cluster *Diagnóstico Dual*.

A prevalência dos doentes com os 3 diagnósticos do cluster *Misto Ansiedade-Depressão* era de 0.01077 e a prevalência dos doentes com as 4 situações do cluster *Diagnóstico dual* era de 0.00063.

A forte associação entre as doenças psiquiátricas não é de estranhar atendendo ás altas taxas de comorbilidade psiquiátrica que têm sido demonstradas na literatura [3, 62].

Os autores discutem posteriormente aspectos relacionados com a técnica da Análise de Clusters, nomeadamente os que se prendem com o algoritmo de clustering, as medidas de proximidade, os critérios para determinar o número e a qualidade dos clusters, bem como a replicabilidade da solução encontrada.

A finalizar os autores concluem que, "a análise de clusters é útil para organizar a investigação, identificando doenças que se agrupam em multimorbilidades que façam sentido.

Embora esta investigação tenha sido conduzida em doentes de cuidados primários, os resultados podem ser generalizados a doentes mais idosos em populações demograficamente semelhantes.

Além disso esta metodologia pode ser usada para informar qualquer sistema de saúde dos padrões de multimorbilidade que ocorrem nas populações por si abrangidas.

Os sistemas que desenvolvam estratégias de gestão de doença baseada em clusters habituais, mais do que em doenças individuais podem melhorar os resultados baseados nos doentes e a sua eficiência" [14].

3.2.2 Patients with multimorbidity in the hospital setting

O objectivo deste estudo, segundo o autor Miera [23] é "estimar a prevalência e descrever as principais características demográficas e de gestão associadas á multimorbilidade em doentes atendidos num hospital universitário".

O estudo decorreu no Hospital Universitário Marques de Valdecilla em Santander, Cantábria, Espanha.

Foram analisados 4310 indivíduos (12.4%) de todas as altas.

Os doentes foram classificados em 7 categorias clínicas, definidas de acordo com a definição funcional do Processo Assistencial Integrado de Atencion al PPP de la Consejeria de Sanidad de la Junta de Andaluzia.

Estes grupos, definidos em listagem anexa ao artigo, abrangem doenças da cardiologia,. osteoarticulares, pneumológicas, hepáticas e inflamatórias intestinais, neurológicas, arteriopatias, diabetes, doenças hematológicas e oncológicas, codificadas de acordo com a ICD-9.

Considerou-se multimorbilidade caso o doente apresentasse 2 das sete categorias.

Analisou-se em cada grupo as variáveis demográficas (idade, sexo, etc.) e as variáveis de gestão (tipo de ingresso (urgente, normal), área de ingresso, serviço responsável pela alta, e tipo de alta (domicilio, outro), bem como o numero de dias de internamento.

O autor verificou que 16.9% (IC a 95% 15.8-18.1) de todas as altas cumpriam o critério de multimorbilidade. Destes 27.4% (IC 24.2-ti0.6) cumpria mais de 2 categorias da definição funcional.

Os resultados estão resumidos nos quadros 3.2 e 3.3.

De realçar que todas as diferenças são estatisticamente significativas ($p<0.001$).

Da leitura destas tabelas o autor conclui que os doentes que apresentam multimorbilidade:

- São mais velhos;
- Predomina o sexo masculino;
- São mais frequentemente admitidos através do serviço de urgência;
- Predominam nas áreas das especialidades médicas;
- Têm um grau de mortalidade intrahospitalar maior;

Quadro 3.2: Caracterização das variáveis com multimorbilidade

Multimorbilidade		
Variáveis	% ou Média	IC (a 95%)
Idade (media)	72.3	71.4-73.2
Sexo (%)		
Masculino	65.1	61.6-68.5
Feminino	34.9	31.5-38.4
Tipo de Ingresso (%)		
Urgência	73.4	70.2-76.6
Resto	26.6	23.4-29.8
Área de Ingresso (%)		
Esp. Medicas	76.4	73.4-79.5
Resto	23.6	20.5-26.6
Tipo de Alta (%)		
Domicilio	84.1	81.5-86.8
Resto	15.9	13.2-18.5
Dias de estadia (Media)	13.3	12.3-14.3

Quadro 3.3: Caracterização das variáveis sem multimorbilidade

Sem Multimorbilidade		
Variáveis	% ou Média	IC (a 95%)
Idade (media)	51.1	50.3-51.9
Sexo (%)		
Masculino	44.1	42.5-45.7
Feminino	55.9	54.3-57.6
Tipo de Ingresso (%)		
Urgência	60	58.4-1.6
Resto	40	38.4-41.6
Área de Ingresso (%)		
Esp. Medicas	34.2	32.6-35.7
Resto	65.8	64.3-67.4
Tipo de Alta (%)		
Domicilio	95.4	94.7-96.1
Resto	4.6	4.0-5.ti
Dias de estadia (Media)	7.6	7.3-7.9

- Têm menos altas para o domicílio;
- Geram maior número de dias de internamento

Este estudo, apesar do seu carácter descritivo, no que difere, dos outros analisados, lança alguma luz sobre a importância da multimorbilidade nos doentes hospitalares.

3.2.3 Multimorbidity Patterns in the elderly: A New Approach of disease Clustering Identifies Complex Interrelations between Chronic conditions

Este estudo, realizado por Schaffer et. all [63] em dados de ambulatório de uma companhia seguradora alemã, a Gmunder ResatzKasse envolveu 63104 homens e 86176 mulheres no grupo etário de 65 e mais anos.

Foram definidos 46 grupos diagnósticos, de acordo com as doenças mais frequentes em Medicina Geral e Familiar, conforme definidas num inquérito (o "ADT Pannel") do Central Research Insitute of Statuory Ambulatory Health Care in Germany.

A forma como foi conduzido o procedimento vem exaustivamente descrita no artigo de Hendrik van der Bussche et. all [70]

Neste caso foi utilizada uma outra técnica de análise multipartidária, concretamente a Análise Factorial.

Esta técnica encontra-se explicada detalhadamente em todos os livros de estatística e, é do conhecimento geral, pelo que não nos iremos deter aqui no seu desenvolvimento. Adiante lhe dedicaremos algum espaço mais diferenciado.

Para o cálculo da prevalência de padrões de multimorbilidade, os autores alocaram os doentes a um padrão se eles tivessem diagnósticos em pelo menos 3 grupos com pesos *factor loadings* de 0.25 no padrão correspondente.

A análise foi feita separadamente para homens e mulheres.

Foram encontrados 3 padrões de multimorbilidade como se mostra no quadro 3.4.

A medida de Kaiser-Meyer-Olkin foi utilizada como indicador da adequação da amostra e, apresentou bons resultados em ambos os sexos.

Os autores discorrem sobre os méritos desta abordagem, em que foram talvez os primeiros na literatura a usá-la, sobre a abordagem da análise de clusters, devido á existência de sobreposições entre os vários padrões

Quadro 3.4: Resultados da Análise Factorial

Prevalência (%)		
Padrões	M	F
Cardiovascular/Metabólico	39	30
Ansiedade/Depressão/Somatização/Dor	22	34
Alterações Neuropsiquiátricas	0.8	6
K-M-O	0.84	0.85

de multimorbilidade, facto que poderá passar despercebido se se utilizar a abordagem por análise de clusters.

Os autores adiantam ainda uma comparação crítica entre vários estudos, que vamos introduzir, á guisa de síntese, uma vez que o estudo que iremos analisar de seguida se orienta numa direcção diferente.

Assim, os autores afirmam que as diferenças entre os vários estudos são significativas devido a:

- Os estudos diferem nas fontes de dados (i.e. administrativos [14], dados de inquéritos [38], ou dados da clínica [50]
- Diferem nas populações (i.e U.S Veteranos [14], Índios Americanos idosos [38]
- Diferem no numero e tipo de grupos de diagnóstico [14, 38, 50]

Apesar destas diferenças na abordagem, existem alguns resultados comuns nestes estudos.

Todos eles reportam um cluster similar cardiovascular, associado com outras doenças, nuns casos doenças metabólicas [14,63], noutros casos acompanhando o AVC [38], o que também, se verificou no estudo presente em análise.

Um outro grupo encontrado foi o cluster da ansiedade/depressão [14], mas sem as alterações somáticas que os autores do presente trabalho relatam.

De uma forma geral os estudos estão todos de acordo, no essencial; existem pequenas discrepâncias, as quais poderão ser motivadas pela metodologia utilizada no estudo presente.

A técnica de exploração através do método de Análise de Clusters parece, desta maneira, uma boa via para encontrar um caminho no labirinto da multimorbilidade de que os autores falam [63]

3.2.4 Identifying Subgroups of Complex Patients with Cluster Analysis

Este último trabalho de Newcomer et all [54] leva-nos para uma abordagem diferente do problema que temos vindo a estudar.

Aqui o pretendido é encontrar, não doenças que se agrupem entre si, mas grupos de doentes que apresentassem 2 ou mais de 17 doenças crónicas frequentes e, que estivessem categorizados no topo dos 20% dos custos totais em cuidados de saúde, num período consecutivo de 2 anos.

Os doentes vieram de uma organização de cuidados de saúde a Kaiser Permanente Colorado (KPCO) que serviu cerca de 430000 membros durante dois anos (2006/2007).

Os investigadores seleccionaram 17 doenças, baseados na prevalência na população geral na prevalência na sua coorte (de 15480 doentes), numa pesquisa na literatura de condições susceptíveis de provocar hospitalização, ou fizessem prever acontecimentos adversos para a saúde dos utentes.

Dessa pesquisa conjunta foi elaborada a seguinte lista de patologias:

1. Diabetes
2. Doença Pulmonar Obstrutiva Crónica (DPOC)
3. Doença Renal Crónica
4. AVC
5. Obesidade
6. Demência
7. Quedas
8. Fractura do colo do fémur
9. Dor Crónica
10. Úlcera Crónica da Pele
11. Cirurgia Ortopédica
12. Cirurgia da Coluna
13. Cirurgia Abdominal
14. Hemorragia Gastrointestinal

15. Cancro (excluindo o cancro da pele não melanoma)

16. Doença Cardíaca (incluindo a doença coronária aguda e a insuficiência cardíaca)

17. Situações do foro psiquiátrico (depressão primária, ansiedade generalizada e doença bipolar)

A partir daqui os autores converteram os diagnósticos em variáveis binárias (0/1) correspondendo á presença ou ausência do diagnóstico.

A base de dados foi dividida em duas folhas aleatoriamente seleccionadas e ambas foram convertidas em matrizes de dissemelhança utilizando o coeficiente de Jacquard.

O processo aglomerativo foi conduzido usando o algoritmo de Ward em cada uma das matrizes.

Os autores encontraram 10 *clusters*, sendo que alguns deles eram constituídos por doentes com multimorbilidades conhecidas, como a diabetes e a obesidade, doença cardíaca e obesidade, doença renal e diabetes, e várias doenças e condições usuais em idosos fragilizados.

Foram igualmente detectados outros *clusters* menos comuns, tais como cirurgia abdominal e ortopédica com a obesidade;

Foram encontrados dois *clusters* com indivíduos mais jovens: doença mental e dor crónica, e doença mental e obesidade.

Duas situações isoladas eram altamente prevalentes em todos os grupos: doença mental (principalmente depressão) e obesidade.

Estes diagnósticos estavam presentes em todos os clusters, com prevalências que variaram de 28% a 100% (doença mental), e entre 38% a 100% (obesidade).

3.3 Conclusão

A análise dos trabalhos referidos dá-nos a ideia que a técnica de Análise Factorial é especialmente útil no desvendar relações muito complexas, quer entre variáveis (patologias), quer entre doentes portadores de patologias bem definidas.

Devemos contudo recordar que esta técnica é um método exploratório de classificação, em que diferentes algoritmos podem produzir diferentes resultados.

Para terminar vamos ficar com as palavras dos autores que salientam que no estudo demonstraram "como a análise factorial pode ser utilizada

para identificar grupos homogéneos de doentes complexos a partir de uma população largamente heterogénea.

Em alternativa, é possível utilizar os resultados desta investigação para enfatizar a necessidade premente de um conjunto de condições nos serviços para indivíduos com altos padrões de consumo de cuidados." [54].

Capítulo 4

Objetivos

Os objectivos deste trabalho foram os seguintes:

Encontrar padrões de multimorbilidade nos doentes com alta hospitalar na Região de Lisboa e Vale do Tejo no período de 209 a 2011, e com duração de internamento hospitalar superior a 1 dia e inferior a 365 dias.

Uma vez definidos esses padrões estudar a forma como se repetem nos internamentos e, no período referido.

Avaliar a importância dos vários predictores (demográficos, administrativos, etc.) na ocorrência da multimorbilidade através de um modelo de regressão logística.

Pretendemos assim tentar encontrar grupos de doentes, portadores de quadros nosológicos uniformes, que levantem necessidades de cuidados orientados de certa forma específicos.

Para alcançar este objectivo necessitamos de vários passos intermédios, que no fundo, constituem objectivos intermédios, e que justificam a marcha geral da análise.

Pretende-se assim: encontrar grupos de doentes, portadores de quadros nosológicos uniformes, que levantem necessidades de cuidados orientados, em certa forma específicos. Para alcançar estes objectivos foram necessários vários passos que, no fundo, constituem objectivos intermédios, e que configuraram a marcha geral da análise.

Nomeadamente:

- Encontrar um número de variáveis associadas entre si, cuja interligação faça sentido e, que constituam a base do processo analítico;

- Uma vez estas definidas, procurar doentes que partilhem entre si um mesmo padrão nosológico;

- Finalmente estudar a forma como esses padrões se repetem ao longo dos anos

Estes resultados foram analisados por ano, isoladamente, a fim de observar se existe alguma regularidade nos padrões, considerados de ano para ano.

Não foram considerados os casos com menos de 1 dia de internamento, não porque esses casos não possam apresentar multimorbilidade, mas porque consideramos que os mesmos pertencem mais ao âmbito dos Cuidados de Saúde Primários, os quais se encontram fora do objectivo deste estudo.

Capítulo 5

Material e Métodos

Neste capítulo apresentamos o material utilizado para o estudo, bem como abordaremos os aspectos teóricos dos métodos utilizados.

Apresentaremos a marcha geral da análise, discutindo mais em pormenor á luz da teoria aqui apresentada, as opções tomadas e as razões das mesmas.

5.1 Material

As bases de dados utilizadas no nosso trabalho são as dos GDH, respeitante aos anos de 2009 a 2011.

De acordo com a definição do Portal da Saúde [1] "os Grupos de Diagnóstico Homogéneo são um sistema de classificação de doentes internados em hospitais de agudos que agrupa doentes em grupos clinicamente coerentes e similares do ponto de vista de consumo de recursos. Corresponde á tradução portuguesa para Diagnosis Related Groups (DRG).Permite definir operacionalmente os produtos de um hospital, que mais não são que o conjunto de bens e serviços que cada doente recebe em função das suas necessidades e da patologia que o levou ao internamento e como parte do processo de tratamento definido".

Os GDH têm já uma longa historia entre nós que remonta a 1984.

Nesse ano iniciou-se um projecto no Ministério da Saúde, destinado a estudar a viabilidade da implementação do sistema em Portugal, á semelhança dos DRG desenhados no final da década de 60, início dos anos 70 por Robert B. Fetter da Universidade de Yale, que aliás viria a integrar como consultor da empresa Health Systems Management Group a equipa que inicialmente pôs mãos á obra.

Em 1989 foram efectuados os primeiros teste de utilização de GDH como base de financiamento do internamento hospitalar.

Com a circular normativa 1/89 do Gabinete do Sr. Secretário de Estado da Saúde a classificação de doentes em GDH generalizou-se e tornou-se obrigatória, sendo as primeiras tabelas de preços GDH a praticar pelo SNS aprovadas pela Portaria nº 409/90 de 1 de Maio.

Para efeitos de codificação das altas hospitalares em termos de diagnósticos e procedimentos, de forma a possibilitar o agrupamento de episódios em GDH, é utilizada a Internacional Classification of Diseases 9th Revision Clinical Modification ICD-9-CM (classificação de diagnósticos e procedimentos que resulta da adaptação efectuada no EUA da International Classifiction of Diseases 9th Revision, ICD-9 da Organização Mundial de Saúde-OMS).

Esta classificação é utilizada em Portugal desde 1989, sendo os dados presentemente registados na aplicação informática WebGDH.

Mensalmente a informação relativa aos GDH de todos os hospitais do SNS é recolhida de forma a integrar a Base de Dados Nacional de Grupos de Diagnósticos Homogéneos (GDH), sediada na Administração Central do Sistema de Saude, I.P. (ACSS).

Embora criado originalmente nos EUA, os conceitos base foram adaptados e desenvolvidos em inúmeros outros países, constituindo-se como suporte quer do financiamento quer da análise de produção hospitalar [1].

Como se pode depreender desta sucinta introdução, independentemente das "origens" dos GDH, estes servem actualmente, no fundamental, para a avaliação da produção hospitalar e sua remuneração.

Contudo, como entre nós não abundam as bases de dados de cariz principalmente epidemiológico, talvez com excepção dos Registos Oncológicos, acabam por ser uma base de trabalho essencial para uma investigação como a que pretendemos levar a cabo.

5.1.1 A estrutura dos GDH

Nos quadros 5.1, 5.2, 5.3 e 5.4 apresentamos o aspecto geral das variáveis do GDH.

Verificamos que existe um número elevado de variáveis que representam dados identificativos (*ano*, *hospital*, *sexo*, *idade*, etc.), dados sobre a movimentação dos doentes (*data_ entrada, hora_ entrada, serv*, etc.) e outras variáveis, que nos interessam particularmente, e que indicam os diagnósticos dos doentes, de acordo com a exposição da secção anterior.

Estas variáveis (*ddx1-ddx20*) representam o conjunto de diagnósticos que foram codificados e, foi sobre elas que trabalhámos da forma explicitada no

Quadro 5.1: Variáveis do GDH

ano	String	28	0	None	None	28	Left	Nominal	Input	
hosp_id	String	4	0	None	None	4	Left	Nominal	Input	
sexo	Numeric	1	0	None	None	8	Right	Unknown	Input	
data_nasc	String	10	0	None	None	10	Left	Nominal	Input	
idade	Numeric	3	0	None	None	8	Right	Unknown	Input	
distrito	Numeric	2	0	None	None	8	Right	Unknown	Input	
concelho	Numeric	2	0	None	None	8	Right	Unknown	Input	
freguesia	Numeric	2	0	None	None	8	Right	Unknown	Input	
data_entrada	String	10	0	None	None	10	Left	Nominal	Input	
data_saida	String	10	0	None	None	10	Left	Nominal	Input	
hora_entrada	Numeric	5	0	None	None	8	Right	Unknown	Input	
hora_saida	Numeric	5	0	None	None	8	Right	Unknown	Input	
dias_int	Numeric	4	0	None	None	8	Right	Unknown	Input	
hosp_to	String	4	0	None	None	4	Left	Nominal	Input	
hosp_from	String	4	0	None	None	4	Left	Nominal	Input	
serv1	String	7	0	None	None	7	Left	Nominal	Input	
ent1	String	10	0	None	None	10	Left	Nominal	Input	
said1	String	10	0	None	None	10	Left	Nominal	Input	
serv2	String	7	0	None	None	7	Left	Nominal	Input	
ent2	String	10	0	None	None	10	Left	Nominal	Input	
said2	String	10	0	None	None	10	Left	Nominal	Input	
serv3	String	7	0	None	None	7	Left	Nominal	Input	
ent3	String	10	0	None	None	10	Left	Nominal	Input	
said3	String	10	0	None	None	10	Left	Nominal	Input	
serv4	String	7	0	None	None	7	Left	Nominal	Input	
ent4	String	10	0	None	None	10	Left	Nominal	Input	
said4	String	10	0	None	None	10	Left	Nominal	Input	
serv5	String	7	0	None	None	7	Left	Nominal	Input	
ent5	String	10	0	None	None	10	Left	Nominal	Input	
said5	String	10	0	None	None	10	Left	Nominal	Input	
serv6	String	7	0	None	None	7	Left	Nominal	Input	
ent6	String	10	0	None	None	10	Left	Nominal	Input	
said6	String	10	0	None	None	10	Left	Nominal	Input	
serv7	String	7	0	None	None	7	Left	Nominal	Input	
ent7	String	10	0	None	None	10	Left	Nominal	Input	
said7	String	10	0	None	None	10	Left	Nominal	Input	

Quadro 5.2: Variáveis do GDH

ent8	String	10	0	None	None	10	Left	Nominal	Input
said8	String	10	0	None	None	10	Left	Nominal	Input
serv9	String	7	0	None	None	7	Left	Nominal	Input
ent9	String	10	0	None	None	10	Left	Nominal	Input
said9	String	10	0	None	None	10	Left	Nominal	Input
serv10	String	7	0	None	None	7	Left	Nominal	Input
ent10	String	10	0	None	None	10	Left	Nominal	Input
said10	String	10	0	None	None	10	Left	Nominal	Input
serv11	Numeric	7	0	None	None	8	Right	Unknown	Input
ent11	String	10	0	None	None	10	Left	Nominal	Input
said11	String	10	0	None	None	10	Left	Nominal	Input
serv12	Numeric	7	0	None	None	8	Right	Unknown	Input
ent12	String	10	0	None	None	10	Left	Nominal	Input
said12	String	10	0	None	None	10	Left	Nominal	Input
serv13	Numeric	7	0	None	None	8	Right	Unknown	Input
ent13	String	10	0	None	None	10	Left	Nominal	Input
said13	String	10	0	None	None	10	Left	Nominal	Input
serv14	Numeric	7	0	None	None	8	Right	Unknown	Input
ent14	String	10	0	None	None	10	Left	Nominal	Input
said14	String	10	0	None	None	10	Left	Nominal	Input
serv15	Numeric	7	0	None	None	8	Right	Unknown	Input
ent15	String	10	0	None	None	10	Left	Nominal	Input
said15	String	10	0	None	None	10	Left	Nominal	Input
serv16	Numeric	7	0	None	None	8	Right	Unknown	Input
ent16	String	10	0	None	None	10	Left	Nominal	Input
said16	String	10	0	None	None	10	Left	Nominal	Input
serv17	Numeric	7	0	None	None	8	Right	Unknown	Input
ent17	String	10	0	None	None	10	Left	Nominal	Input
said17	String	10	0	None	None	10	Left	Nominal	Input
serv18	Numeric	5	0	None	None	8	Right	Unknown	Input
ent18	String	10	0	None	None	10	Left	Nominal	Input
said18	String	10	0	None	None	10	Left	Nominal	Input
serv19	Numeric	5	0	None	None	8	Right	Unknown	Input
ent19	String	10	0	None	None	10	Left	Nominal	Input
said19	String	10	0	None	None	10	Left	Nominal	Input
serv20	Numeric	5	0	None	None	8	Right	Unknown	Input

Quadro 5.3: Variáveis do GDH

said20	String	10	0	None	None	10	Left	Nominal	Input	
ddx1	String	5	0	None	None	5	Left	Nominal	Input	
ddx2	String	5	0	None	None	5	Left	Nominal	Input	
ddx3	String	5	0	None	None	5	Left	Nominal	Input	
ddx4	String	5	0	None	None	5	Left	Nominal	Input	
ddx5	String	5	0	None	None	5	Left	Nominal	Input	
ddx6	String	5	0	None	None	5	Left	Nominal	Input	
ddx7	String	5	0	None	None	5	Left	Nominal	Input	
ddx8	String	5	0	None	None	5	Left	Nominal	Input	
ddx9	String	5	0	None	None	5	Left	Nominal	Input	
ddx10	String	5	0	None	None	5	Left	Nominal	Input	
ddx11	String	5	0	None	None	5	Left	Nominal	Input	
ddx12	String	5	0	None	None	5	Left	Nominal	Input	
ddx13	String	5	0	None	None	5	Left	Nominal	Input	
ddx14	String	5	0	None	None	5	Left	Nominal	Input	
ddx15	String	5	0	None	None	5	Left	Nominal	Input	
ddx16	String	5	0	None	None	5	Left	Nominal	Input	
ddx17	String	5	0	None	None	5	Left	Nominal	Input	
ddx18	String	5	0	None	None	5	Left	Nominal	Input	
ddx19	String	5	0	None	None	5	Left	Nominal	Input	
ddx20	String	5	0	None	None	5	Left	Nominal	Input	
cext1	String	5	0	None	None	5	Left	Nominal	Input	
cext2	String	5	0	None	None	5	Left	Nominal	Input	
cext3	String	5	0	None	None	5	Left	Nominal	Input	
cext4	String	5	0	None	None	5	Left	Nominal	Input	
cext5	String	5	0	None	None	5	Left	Nominal	Input	
proc1	Numeric	4	0	None	None	8	Right	Unknown	Input	
proc2	Numeric	4	0	None	None	8	Right	Unknown	Input	
proc3	Numeric	4	0	None	None	8	Right	Unknown	Input	
proc4	Numeric	4	0	None	None	8	Right	Unknown	Input	
proc5	Numeric	4	0	None	None	8	Right	Unknown	Input	
proc6	Numeric	4	0	None	None	8	Right	Unknown	Input	
proc7	Numeric	4	0	None	None	8	Right	Unknown	Input	
proc8	Numeric	4	0	None	None	8	Right	Unknown	Input	
proc9	Numeric	4	0	None	None	8	Right	Unknown	Input	
proc10	Numeric	4	0	None	None	8	Right	Unknown	Input	

capítulo seguinte (Marcha Geral da Análise).

5.2 Métodos

Nesta secção iremos fazer uma descrição sucinta da fundamentação teórica dos vários métodos que utilizaremos no tratamento dos dados.

Para além dos aspectos meramente descritivos, que não merecem nenhuma consideração em especial, iremos apresentar o *rationale* das duas técnicas utilizadas: Análise Factorial e Regressão Logística .

Os aspectos peculiares do nosso caso, bem como as opções escolhidas e as decisões tomadas, serão explicadas em pormenor mais á frente.

5.2.1 Análise Factorial

Introdução

Nos últimos anos, a Análise Factorial tornou-se acessível a um grupo alargado de investigadores e estudantes, devido sobretudo ao desenvolvimento de computadores de alto desempenho, bem como do respectivo "software".

Como consequência, foi possível que pessoas com formação matemática não muito desenvolvida pudessem, mesmo assim, explorar o potencial do método para uso da sua própria investigação [41].

A Análise Factorial tem sido amplamente utilizada, da economia à sociologia da psicologia às ciências da saúde.

Na economia tem sido utilizada para derivar um conjunto de variáveis não correlacionadas, para futura análise, quando o uso de variáveis fortemente correlacionadas pode conduzir a resultados enganadores na análise da regressão.

As ciências políticas usam-na para, por exemplo, comparar os atributos das nações no respeitante a uma multiplicidade de variáveis políticas e sócio-económicas, num esforço para determinar quais as características mais importantes na classificação das nações (saúde, dimensão, etc.) (op. citada); os sociólogos, por exemplo, tentaram encontrar "grupos amigáveis" verificando quais as pessoas que mais se relacionam umas com as outras (e igualmente as que não se relacionam entres si): psicólogos e educadores têm usado esta técnica a fim de determinar a forma como as pessoas percebem diferentes "estímulos", e os categorizam em respostas diferentes (como se interrelacionam por exemplo os diferentes elementos da linguagem) (op. citada).

Quadro 5.4: Variáveis do GDH

proc11	Numeric	4	0		None	None	8	Right	Unknown	Input
proc12	Numeric	4	0		None	None	8	Right	Unknown	Input
proc13	Numeric	4	0		None	None	8	Right	Unknown	Input
proc14	Numeric	4	0		None	None	8	Right	Unknown	Input
proc15	Numeric	4	0		None	None	8	Right	Unknown	Input
proc16	Numeric	4	0		None	None	8	Right	Unknown	Input
proc17	Numeric	4	0		None	None	8	Right	Unknown	Input
proc18	Numeric	4	0		None	None	8	Right	Unknown	Input
proc19	Numeric	4	0		None	None	8	Right	Unknown	Input
proc20	Numeric	4	0		None	None	8	Right	Unknown	Input
morf_tum1	Numeric	5	0		None	None	8	Right	Unknown	Input
morf_tum2	Numeric	5	0		None	None	8	Right	Unknown	Input
morf_tum3	Numeric	5	0		None	None	8	Right	Unknown	Input
morf_tum4	Numeric	15	0		None	None	8	Right	Unknown	Input
morf_tum5	Numeric	15	0		None	None	8	Right	Unknown	Input
dsp	Numeric	2	0		None	None	8	Right	Unknown	Input
peso_nasc	Numeric	4	0		None	None	8	Right	Unknown	Input
adm_tip	Numeric	1	0		None	None	8	Right	Unknown	Input
gdh	Numeric	3	0		None	None	8	Right	Unknown	Input
gcd	Numeric	2	0		None	None	8	Right	Unknown	Input
agr_versao	String	4	0		None	None	4	Left	Nominal	Input
portaria	Numeric	2	0		None	None	8	Right	Unknown	Input
tipo_gdh	String	1	0		None	None	1	Left	Nominal	Input
interv_cir	String	10	0		None	None	10	Left	Nominal	Input
mot_transf	Numeric	2	0		None	None	8	Right	Unknown	Input
data_urgencia	String	10	0		None	None	10	Left	Nominal	Input
hora_urgencia	Numeric	5	0		None	None	8	Right	Unknown	Input
tipo_port	String	3	0		None	None	3	Left	Nominal	Input
sns	Numeric	1	0		None	None	8	Right	Unknown	Input
doente_eq	Numeric	5	0		None	None	8	Right	Unknown	Input
n_ficticio_utente	Numeric	9	0		None	None	8	Right	Unknown	Input

As ciências da saúde têm utilizado abundantemente esta técnica, citando como exemplo o estudo de Chaitman et al. [12], referido por Fisher [68], sobre os efeitos da cirurgia de "by-pass" coronário na sobrevivência em grupos de doentes com doença coronária afectando a coronária esquerda..

Esta técnica, conforme explicamos na secção seguinte, visa representar um conjunto de variáveis, em termos de um número mais reduzido de variáveis hipotéticas - os factores -, com as quais se pretende explicar o fenómeno em estudo.

Num extremo, o investigador pode não ter qualquer ideia acerca da real dimensão dos factores explicativos dos dados observados.

Neste caso a Análise Factorial pode ser usada como um meio para extrair da amostra um número mínimo de factores hipotéticos que podem explicar a covariância observada e, como um meio de explorar os dados com vista à redução dos mesmos.

Esta forma de uso é exploratória, sendo que a maioria das aplicações das ciências sociais, nomeadamente no nosso caso, seguem este caminho.

Porém o uso da Análise Factorial não necessita de se confinar à exploração da dimensão subjacente dos dados.

Dependendo dos conhecimentos do investigador, o método pode ser usado como um meio de testar hipóteses específicas; neste caso considera-se a análise como confirmatória (Reis Elizabete [21]).

Por exemplo, o investigador pode antecipar, ou inferir, da existência de duas dimensões diferentes e, que certas variáveis pertencem á primeira dimensão, enquanto outras pertencem à segunda.

Se a Análise Factorial é utilizada dentro desta perspectiva, então ela é usada como meio de confirmação de uma dada hipótese e, não só como meio de explorar a dimensão subjacente.

Trata-se assim como atrás se disse de Análise Factorial confirmatória.

A aplicação do método pode não ser, em certas circunstâncias, tão simples como se pode depreender das afirmações anteriores.

Em certos casos, a divisão entre as perspectivas exploratória e confirmatória não se torna tão evidente como isso.

É importante reforçar aqui a ideia que a Análise Factorial é uma técnica extremamente promissora no campo das ciências sociais e humanas, e que o seu uso não requer profundos conhecimentos teóricos da técnica estatística, embora necessite, isso sim, da compreensão dos seus fundamentos lógicos e conceptuais.

Uma última palavra, a finalizar esta breve introdução, para a importância diríamos mesmo, para a imprescindibilidade, do uso dos computadores na execução desta técnica.

De facto, a não ser em casos muito pontuais, com dados absolutamente distantes da realidade prática, nada se pode levar a cabo sem o uso desse precioso auxiliar.

Por outro lado, existe uma multiplicidade de "software" estatístico apto a desempenhar as suas funções neste campo.

Programas de índole geral, como o SAS ou o SPSS, contém em si os módulos necessários á execução das operações adiante descritas.

Neste nosso trabalho, contudo, utilizaremos um outro programa o STATA 12.0, um software que embora pouco divulgado em Portugal, nos meios empresariais, é amplamente utilizado em todo o mundo, nomeadamente nos EUA, sobretudo na área da epidemiologia, uma vez que para além dos aspetos comuns a todos os programas de uso geral, contém ainda algumas funções que se adaptam especialmente bem a esta área, que é a do nosso interesse particular.

Existem para além disso razões especiais para o seu uso neste caso como adiante explicaremos.

Fundamentos lógicos da análise

A Análise Factorial baseia-se no pressuposto fundamental de que certos factores, em número inferior ao das variáveis observadas, são responsáveis pela covariância entre elas.

A base teórica deste método está particularmente bem explicada na obra clássica de Gorsuch [26], bem como nos textos já citados de Kim and Mueller [40, 41].

Todos os livros de estatística, porém, contêm capítulos dedicados ao assunto.

Iremos seguir aqui a linha de exposição de Fisher and van Belle [68] ficando somente nos tópicos principais.

Para melhor esclarecimento imaginemos o modelo mais simples em que um factor subjacente é responsável pela covariância entre duas variáveis observadas.

Tal exemplo pode ser demonstrado pela Figura 5.1.

Como se pode observar X_1 é uma soma ponderada de F e U_1 e X_2 é uma soma ponderada de F e U_2.

Como F é um factor comum a X_1 e X_2 pode ser chamado **factor comum**.

Igualmente, porque U_1 e U_2 são únicos para cada variável observada são definidos como **factores únicos**.

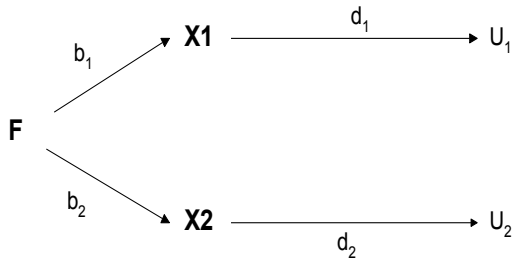


Figura 5.1: Modelo com duas variáveis e um fator comum

Daqui podemos derivar as seguintes equações:

$$X_1 = b_1F + d_1U_1 \quad (5.1)$$
$$X_2 = b_2F + d_2U_2 \quad (5.2)$$

Para além do mais o diagrama indica igualmente que não existe covariância entre F e U_1, F e U_2 ou entre U_1 e $U2_2$.

Logo:

$$Cov(F, U_1) = Cov(F, U_2) = Cov(U_1, U_2) = 0 \quad (5.3)$$

As três equações 5.1, 5.2 e 5.3 definem um sistema linear de **Análise Factorial** como o apresentado na Figura 5.1.

Se considerarmos queX_1 e X_2 são as variáveis observadas, e que F, U_1 e U_2 são as variáveis não observadas, temos então o modelo de factores comuns mais simples.

Note-se que existe um maior número de factores (F, U_1,U_2) do que de variáveis observadas (X_1, X_2); contudo somente o factor F é comum a ambas as variáveis e portanto o número de factores comuns é inferior ao número de variáveis.

Finalmente recordemos que as variáveis observadas (X_1, X_2) são criadas a partir de variáveis desconhecidas e inobserváveis (F, U_1, U_2); a estas variáveis que são a origem das variáveis observadas chamamos **variáveis hipotéticas** ou **factores hipotéticos**.

Destes os que estão envolvidos na criação de mais do que uma variável observada são chamados **factores comuns**; os que são usados para criar somente uma variável observada são chamados **factores únicos**.

Se analisarmos o problema de outro ângulo, vemos que é possível e, eventualmente mais útil para nós, definir o modelo em função das variáveis observadas X_1 e X_2.

$$X_1 = E[X_1] + \lambda_{11}F_1 + \lambda_{12}F_2 + \cdots + \lambda_{1k}F_k + e_1 \quad (5.4)$$
$$X_p = E[X_p] + \lambda_{p1}F_1 + \lambda_{p2}F_2 + \cdots + \lambda_{pk}F_k + e_p \quad (5.5)$$

Desta forma as equações 5.4 e 5.5 apresentam o modelo de uma forma mais geral, considerando mais do que um factor comum.

Assim, vemos que neste modelo cada X é uma soma linear de vários factores, mais alguma variabilidade residual remanescente.

Neste modelo cada X é assim igual ao seu valor esperado mais uma soma linear de k factores mais um termo para a variabilidade residual.

Devemos então salientar alguns aspectos e pressuposto do modelo, a saber:

- Como acima dissemos os factores F_j não são observados; somente os $X_1....X_p$ são observados, embora as variáveis X_i sejam expressas em termos do menor número de factores F_j;
- Os e_i (que também não são observados) representam a variabilidade em X_i não explicada pelos factores comuns F_j. Não assumimos que estes termos de variabilidade residual tenham a mesma distribuição;
- Habitualmente o número de k factores é desconhecido, e é determinado a partir dos dados.

Os pressupostos do modelo, para além das equações lineares acima definidas são os seguintes:

- Os factores F_j são padronizados, isto é, têm média 0 e variância 1;
- Os factores F_j não estão correlacionados entre si e, não estão correlacionados com os termos e_i. (Podem existir modelos em que este pressuposto não seja cumprido, nomeadamente a correlação entre os factores F_j).
- Os e_i não estão correlacionados entre si e com os F_j e têm média 0 e podem ter diferentes variâncias.

Seja ψ_i a variância de e_i.

Nos pressupostos do modelo, acima definidos, a variância de cada X_i pode ser expressa em termos dos coeficientes λ_{ij} dos factores e da variância residual ψ_i.

A equação 5.6 apresenta a relação para os k factores:

$$var\,(X_i) = \lambda_{i1}^2 + \cdots + \lambda_{ik}^2 + \psi_i \tag{5.6}$$

Assim a variância de cada X_i é a soma dos quadrados dos coeficientes dos factores, mais a variância de e_i.

A variância de $X_i$tem assim dois componentes:

A soma dos coeficientes λ_{ij} ao quadrado depende dos factores; os factores contribuem em comum para todos os X_i´s.

Os e_i´s correlacionam-se somente com a sua variável X_i e não com quaisquer outras variáveis no modelo.

Dividimos assim a variância numa parte relacionada com com os factores que cada variável tem em comum, e uma parte única relacionada com a variabilidade residual.

Isto conduz-nos á seguinte definição:

Definição: A grandeza $c_i = \sum_{j=1}^{k} \lambda_{ij}^2$ denomina-se parte comum da variância de X_i, c_i também designada por **comunalidade** de X_i, ψ_i denomina-se parte única ou específica da variância de X_i

Embora a Análise Factorial se destine a explicar a relação entre as variáveis, e não a variância das variáveis individuais, se a comunalidade for grande comparada com os valores específicos das variáveis, então o modelo tem igualmente sucesso na explicação, não só das relações entre as variáveis, mas igualmente a variabilidade em termos de factores comuns.

Um outro conceito necessário para a compreensão e interpretação do modelo é dado na seguinte definição.

Definição: Os coeficientes λ_{ij} são denominados de "factor loadings" ou mais simplesmente *loadings*, correspondem aos peso na variável X_i no fator F_j, na prática ao valor b_i da Figura 5.1

Correspondem á covariância entre X_i e F_j.

Munidos destes conceitos, podemos rapidamente analisar alguns aspectos complementares da análise factorial.

Indeterminação do espaço vetorial-Rotações

A estimação dos coeficientes das variáveis que nem sequer são observadas pode parecer um pouco estranho.

É um pouco difícil imaginar que se possa estimar todas elas.

De facto, não só é possível estimar F como igualmente estimá-lo até um certo limite de indeterminação.

É necessário definir essa indeterminação em termos matemáticos [68].

Matematicamente os factores serão únicos, com excepção para possíveis combinações lineares.

Geometricamente podemos pensar os factores, como por exemplo num caso com $K = 2$, como correspondendo a valores num plano.

Imaginemos esse plano num espaço tridimensional.

Por exemplo, o sub-espaço correspondente aos dois factores, i. e. ao plano poderá ser a folha de papel em que desenhamos.

Dentro deste espaço tridimensional a ANÁLISE FACTORIAL determinará que plano contém os dois factores.

Contudo, quaisquer duas dimensões perpendiculares no plano dos factores corresponderá a factores que igualmente ajustam bem os dados em termos de explicação das covariâncias ou correlações entre as variáveis.

Assim temos os factores identificados, até um certo ponto, mas temos a liberdade de os rodar dentro de um sub-espaço.

Esta indeterminação permite-nos, como se disse, jogar com diferentes combinações de factores, i.e. rotações, de forma que os factores se possam considerar "fáceis de interpretar".

Considerem-se a figura 5.2.

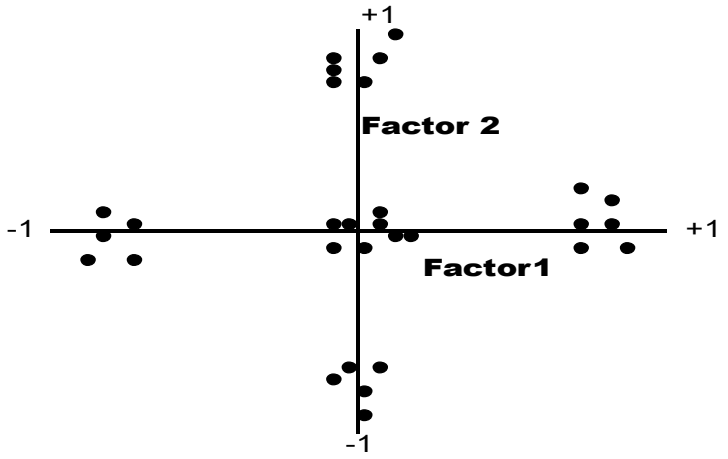


Figura 5.2: Padrão de loadings de dois factores

Esta figura representa um gráfico de pesos dos factores 1 e 2, com cada ponto pertencente à variável X_i.

Nesta figura observa-se um padrão muito regular.

As variáveis correspondentes aos pontos no eixo do fator 1 (+-1) ou no fator 2 (+-1) são variáveis associadas aos dois factores.

As variáveis grafadas junto a 0 têm pouco a ver com os dois factores. Na figura 5.3 observamos uma situação bastante diferente.

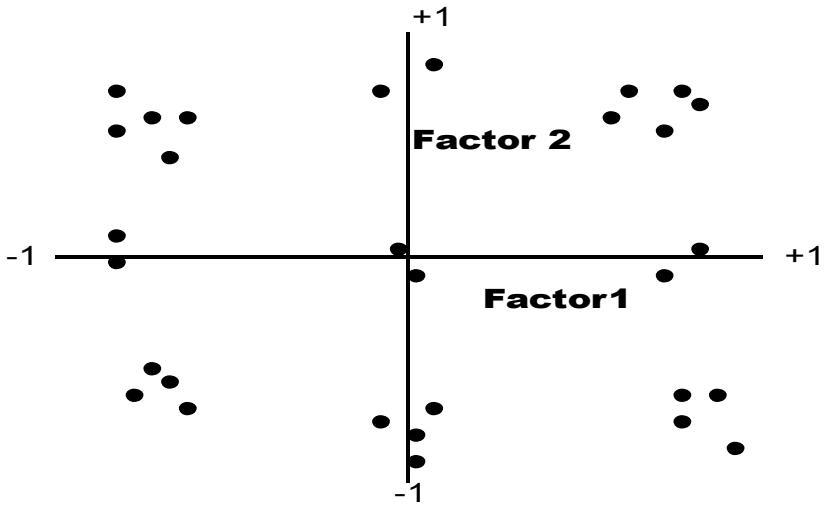


Figura 5.3: Outro Padrão

Repare-se que aqui se torna mais difícil interpretar a distribuição dos pontos, em termos dos factores.

Contudo reparemos na figura 5.4

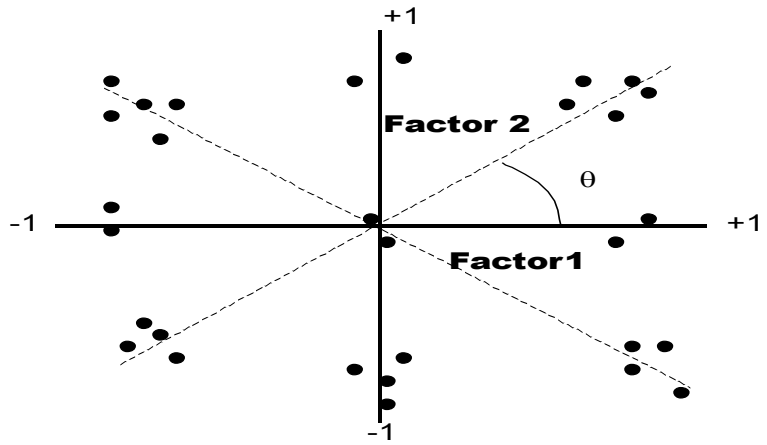


Figura 5.4: Rotação ortogonal

Se procedermos à rotação dos eixos por θ conforme indicado, temos de novo factores de fácil interpretação, i.e. cada factor fica associado a um subconjunto das X_i variáveis.

Verificando os pontos, desenhando as linhas e decidindo qual o valor do ângulo θ, procedemos então a uma rotação visual.

Esta poderá ser possível num espaço com $k <= 2$ variáveis; contudo imagine-se o que se passará com um espaço $k > 2$.

Torna-se impossível na prática.

Percebe-se assim o interesse no desenvolvimento de um algoritmo que permita decidir se uma dada rotação, para todos ou qualquer dos factores, é desejável.

O software do computador, encontrará então a melhor rotação.

O método mais popular de rotação é o método proposto por Kaiser (1958), como citado por Gorsuch [26] denominado *varimax*.

Trata-se de um método ortogonal e pretende que para o componente principal existem apenas alguns pesos significativos e tudo o resto seja próximo de θ [21].

Este usa a ideia de maximizar a soma das variâncias dos quadrados dos pesos dos factores.

Note-se que as variâncias são elevadas quando b_{ij}^2 está próximo de 0 ou 1 em qualquer das colunas.

A fim das variáveis com uma grande comunalidade não serem sobrevalorizadas, são usados valores ponderados.

Suprimimos aqui as fórmulas que podem contudo ser consultadas nos textos já referidos.

Voltando á rotação visual, suponhamos que temos um padrão como o representado na figura 5.5

Verificamos que não é possível desenhar eixos perpendiculares para os quais os pesos sejam $+1$ ou -1: mas se tomar-mos dois eixos correspondentes às linhas tracejadas a interpretação simplificar-se-á.

Os factores correspondentes às duas linhas tracejadas estão correlacionados entre si e podemos interrogar-nos se de facto se tratará de factores "separados".

Estes factores são denominados factores oblíquos, sendo que o termo oblíquo deriva da figura geométrica e do facto de em geometria linhas obliquas serem as que não se intersectam em ângulo recto.

Existem variados métodos para obter rotações obliquas, com nomes como *Oblimax, Biquartimin, Binormamim e Maxplane*.

Referências a estes métodos podem ser encontrados em Gorsuch [26].

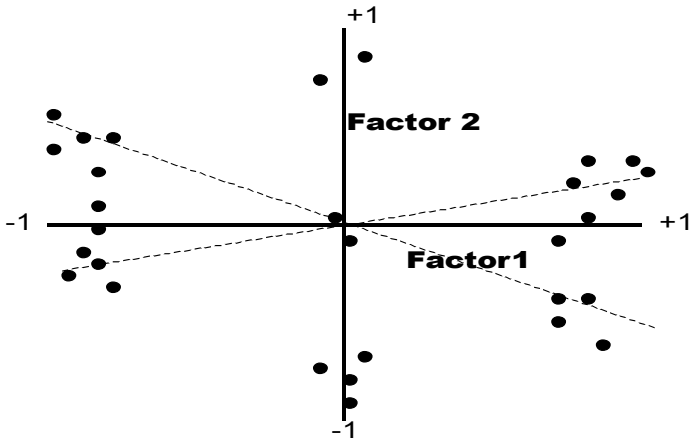


Figura 5.5: Rotação oblíqua

Extração dos factores e sua interpretação

Vários são os métodos descritos para a extração dos factores.

Fazemos uma breve referência a um método muito utilizado e que é o dos *mínimos quadrados*.

O principio por detrás da aproximação dos mínimos quadrados á análise factorial é o de minimizar a correlação residual depois de extrair um certo número de factores e, da avaliação do grau de ajustamento entre a correlação observada e a correlação sob o modelo.

Porque se pode sempre reproduzir a correlação observada imaginando tantos factores quantas as variáveis, e porque o ajustamento melhora quando o número de factores hipotéticos aumenta, o método dos mínimos quadrados (MQ) assume que partimos com a hipótese que K número de factores (K menor que o número de variáveis) são responsáveis pela correlação observada.

O processo utilizado para obter a solução é, em linhas gerais, como segue:

1. Assumimos que K factores são responsáveis pela correlação observada;

2. Obtemos uma estimativa inicial das comunalidades (como se disse o quadrado do coeficiente de correlação múltipla R_2 e as restantes);

3. Extraímos os K factores que melhor reproduzem a matriz de correlação observada (de acordo com o princípio dos mínimos quadrados);

4. A fim de obter o padrão que melhor possa reproduzir a matriz de covariância observada as comunalidades são reestimadas na base do padrão encontrado no passo anterior;

5. O processo é repetido até que não se observem melhorias no modelo.

É este processo "iterativo" que os computadores usam e cujo output iremos, primeiro explicar, e depois observar num exemplo de marcha geral do processo na secção seguinte e, finalmente na interpretação dos dados do nosso estudo.

Número de factores e sua interpretação

A escolha do número de factores que explicam a variância do nosso modelo não é como se vê tarefa fácil.

Felizmente que temos hoje a ajuda preciosa dos computadores nesta tarefa.

Contudo, para além da definição da própria estratégia da análise, torna-se durante a marcha da mesma necessário tomar decisões para as quais o investigador deve estar preparado.

Iremos por isso aqui abordar dois aspetos, a saber:

- Especificação dos valores próprios;

- Critérios de interpretação.

Um dos critérios mais usuais para determinar o número de factores comuns é o de aceitar todos os que tenham valores próprios acima de 1, quando a matriz de correlação (não ajustada) é decomposta.

Este critério estabelecido por Kaiser, leva o seu nome e, apesar de tudo tem demonstrado na prática adaptar-se bastante bem á realidade observada.

O autor demonstra que os factores cujos valores próprios estejam abaixo da unidade contribuem de modo muito modesto para a explicação da variância, pelo que poderão ser rejeitados.

Um outro teste, igualmente usado foi introduzido por Cattell [10] e baseia-se na análise do polígono de valores próprios (*scree plot*).

Este polígono representa a variância dos componentes principais, isto é, os seus valores próprios, como a Figura 5.6 mostra.

Como se pode observar no ponto em que o gráfico inflete no sentido da horizontal, devemos para a extração de factores.

Quer dizer portanto que só deverão ser considerados os factores à esquerda desse ponto.

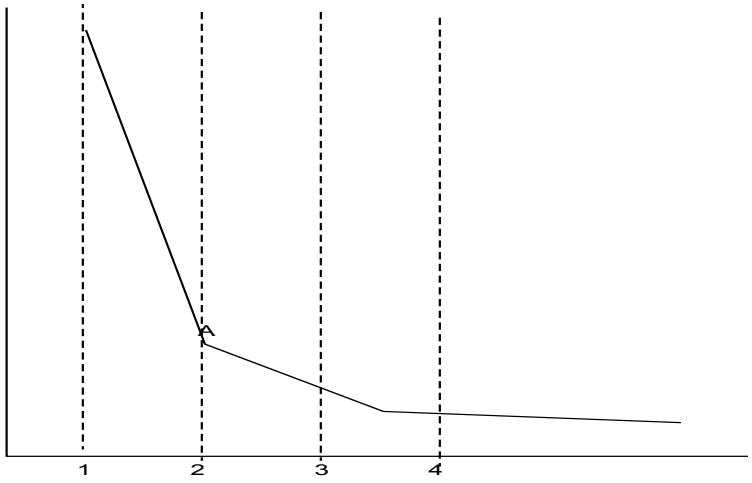


Figura 5.6: Polígono de valores próprios

Para terminar esta secção vamos abordar um aspeto muito importante, que é o da interpretação dos factores.

É tentador, mas perigoso, avançar com interpretações muito pormenorizadas dos factores, sem ter um conhecimento substantivo e largo da matéria em apreço.

Como em tudo na Estatística, os resultados analíticos têm que ser ponderados em função dos conteúdos lógicos e conceptuais, nos quais se tem de verificar com atenção se aqueles encaixam, ou não.

Gorsusch [26] citado por Fisher [68] avança com algumas "regras de ouro" na interpretação dos factores, que nos parece pertinente aqui enfatizar.

- Os factores só devem ser interpretados por indivíduos com exaustivo conhecimento da área em estudo;

- O sumário da interpretação é apresentado com o nome do factor; o nome pode ser somente descritivo, ou pode sugerir uma explicação causal para a ocorrência dos factos; dado que o nome dos factores é aquilo que a maior parte dos leitores do relatório do investigador irá ler, deverá ser cuidadosamente escolhido, ou mesmo, deve até em muitos casos nem sequer ser escolhido;

- A prática muito espalhada de olhar a interpretação global de um factor como confirmada, só porque a análise post-hoc faz sentido, deve ser

banida. A interpretação dos factores só deve ser considerada como uma hipótese para novos trabalhos.

Marcha Geral do processo da Análise Factorial

Para terminarmos esta parte vamos apresentar um exemplo extraído de Hamilton [29] em que se apresenta o caminho geral a seguir num processo de análise factorial, que aliás seguiremos no nosso trabalho.

Vamos aqui abordar as considerações atrás explicadas e, vamos interpretar o "output" do computador a fim de exemplificar a marcha geral.

Esta baseia-se em quatro passos fundamentais:

1. Extracção dos factores iniciais;
2. Rotação para simplificação da estrutura;
3. Decisão sobre quais os factores a extrair;
4. Obter e usar os *scores* dos factores

Não nos podemos esquecer contudo que o processo de estimação é iterativo e que ele deverá ser repetido até obter os melhores resultados.

Portanto os passos anteriores converter-se-ão assim em:

1. Extracção iterativa dos factores;
2. Rotação para simplificação da estrutura;
3. Decisão sobre quais os factores a reter;
4. Avaliação do modelo obtido com repetição de 1 a 3 até obter o melhor resultado;
5. Obter e utilizar os *scores* dos dos factores.

O exemplo que vamos usar é retirado de Blocken e Eckberg [6] citado como se disse por Hamilton [29].

Os inquéritos/questionários para identificar perfis psicológicos, ou atitudes, fornecem habitualmente um bom material para uso da Análise Factorial.

No fundo, sempre que se pretende encontrar um comportamento que se possa admitir como um padrão de actuação, estamos em boas condições para aplicar o método.

Trata-se assim de reduzir um conjunto de variáveis a um número inferior de factores explicativos.

O quadro 5.5 lista seis questões de um inquérito conduzido em Tulsa, Oklahoma (Block e Eckberg, 1989 [6].

Quadro 5.5: Inquérito na área de Tulsa sobre problemas ambientais

Como acha que estes problemas o afectarão	
Z1 (Deepwell)	Infiltração de resíduos químicos em poços subterrâneos
Z2 (Chander)	Incêndios subterrâneos em Chandler Park
Zti (Tornados)	Tornados
Z4 (Floods)	Inundações
Z5 (Airpol)	Poluição aérea
Z6 (Rivpol)	Poluição de cursos de água

Pretende-se a resposta a questões que reflectem a preocupação sobre acontecimentos que possam pôr em perigo o meio ambiente.

Este inquérito, produziu dados em seis variáveis diferentes, mas conceptualmente relacionadas.

Aplicada a estes dados, a Análise Factorial serve dois objectivos:

1. Traduzir a resposta das pessoas a essas seis questões específicas, denotando preocupações gerais como "preocupação sobre a poluição", ou "medo de catástrofes naturais; a Análise Factorial ajuda-nos a identificar e medir essas variáveis latentes;

2. Se as variáveis latentes emergirem plausivelmente, simplificam a análise subsequente; poucas coordenadas de factores (factor scores) podem substituir vários itens individuais.

A Análise Factorial bem sucedida requer um padrão numa matriz de correlação; subgrupos de variáveis que se correlacionam mais fortemente que outras.

O quadro 5.6 dá-nos uma matriz de correlação para as seis variáveis do estudo de Tulsa (n=199)

Estas correlações são moderadas, como é aliás habitual nos inquéritos.

Podemos contudo discernir alguns padrões.

Por exemplo:

Tornado e inundações correlacionam-se entre si moderadamente (r=.4052), mas debilmente com outras variáveis.

Outras correlações moderadas verificam-se entre a infiltração de resíduos químicos e o incêndio de Chanddler (r=.3861).

Quadro 5.6: Matriz de correlação

	z1	z2	zti	z4	z5	z6
z1	1.00	•	•	•	•	•
z2	.472	1.00	•	•	•	•
z3	.113	.202	1.00	•	•	•
z4	.092	.408	.405	1.00	•	•
z5	.280	.166	.152	.071	1.00	•
z6	.336	.258	.100	.151	.386	1.00

A análise de componentes principais dos dados de Tulsa, conduziu ao polígono de valores próprios da figura 5.7

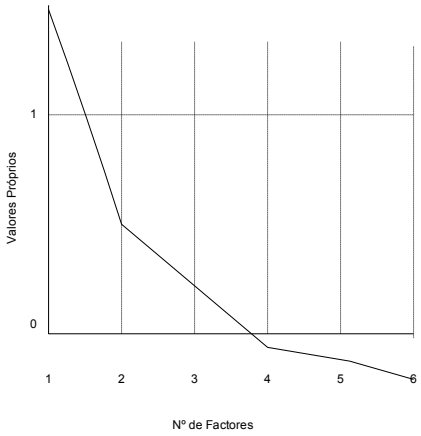


Figura 5.7: Polígono de valores próprios para os factores principais do inquérito de Tulsa

Como podemos observar, apenas três (3) factores têm valores próprios positivos.

Após uma rotação obliqua os três primeiros factores na figura 5.8 fazem sentido.

Após a análise factorial, Blocke e Eckberg examinaram como os scores dos factores variavam no que respeita às características demográficas.

Verificaram que os valores médios dos scores dos factores representando a preocupação com o ambiente local (F3) eram significativamente maiores no sexo feminino que no masculino.

Testando esta nova variável, usando uma "dummy variable" verificou-se que a sua média é bastante maior nos homens que nas mulheres $-.128$

(Componentes principais; 3 factores retidos)

Factor	Valor Próprio	Diferença	Proporção	Cumulativo
1	1.35194	0.88877	0.9929	0.9929
2	0.46317	0.29830	0.3402	1.3331
3	0.16487	0.28236	0.1211	1.4542
4	-0.11749	0.07032	-0.0863	1.3679
5	-0.18781	0.12528	-0.1379	1.2299
6	-0.31309		-0.2299	1.00

(Rotação promax)

Pesos dos factores após rotação

Variável	1	2	3	Uniqueness
deepwell	**0.54661**	-0.02946	0.14547	0.59327
chandler	**0.61054**	0.03497	-0.03720	0.64047
tornados	0.06995	**0.54955**	-0.02505	0.68031
floods	-0.06653	**0.54282**	0.03733	0.71092
airpol	0.04270	0.00781	**0.49339**	0.72551
rivpol	0.13743	0.01013	**0.47524**	0.66954

(Correlações entre factores)

(obs=199)

	F_1	F_2	F_3
F_1	1.000		
F_2	0.4957	1.000	
F_3	0.8732	0.5503	1.00

Figura 5.8: Análise de componentes principais do inquérito de Tulsa

contra 1.66 - $P(0.004)$.

Uma vez definidas os "factor scores" podemos estudá-las como qualquer outra variável, usando todas as técnicas analíticas conhecidas.

Ficam desta forma demonstrados em traços muito largos os fundamentos teóricos da Análise Factorial que iremos utilizar.

Contudo, no nosso caso, será necessário um pequeno ajustamento á teoria, devido ao facto das nossas variáveis serem todas, como adiante se verá, variáveis binárias (0/1).

A Análise Factorial é assim um método particularmente dedicado ao estudo de variáveis contínuas, até porque se baseia na utilização de matrizes de covariância/correlação entre as mesmas.

Daí o facto de se utilizar o coeficiente de correlação de Pearson na construção dessas matrizes.

Ora no nosso caso tal não se torna possível, devido á natureza das variáveis, como se disse. como fazer então?

Alguns autores, nomeadamente portugueses [51], preconizam a utilização do coeficiente V de Cramer, com o qual se construiria uma matriz de correlação posteriormente analisada pelos métodos atrás referidas.

A solução mais amplamente divulgada passa, no entanto, por construir

uma matriz de correlação dita policórica (tetracórica no caso de variáveis dicotómicas).

Façamos uma breve análise deste método seguindo o texto de Uebersax [66].

Correlação policórica e tetracórica

O coeficiente de correlação tetracórico (Pearson 1901), para dados binários, e o coeficiente de correlação policórico para dados ordinais, são excelentes formas de medir a concordância entre as observações.

Estimam o que a correlação entre as variáveis deveria ser se as observações fossem obtidas numa escala contínua; são, teoricamente, invariantes ás modificações no número, ou dimensão das categorias.

Intuitivamente consideremos o exemplo de dois psiquiatras (A e B) que realizam o diagnóstico de presença/ausência de Depressão major. embora o diagnóstico seja dicotómico, aceitamos que a Depressão como patologia tem uma distribuição contínua na população (como poderia ser qualquer outra doença que pretendemos estudar).

Ao diagnosticar um dado caso, um dos médicos considera o nível, Y, de depressão do caso relativo a dado limiar, t; se o nível avaliado for superior a t um diagnóstico positivo (1) é feito; de outra forma o diagnóstico será negativo (0).

Observemos a figura 5.9.

Se entendermos *Y1, Y2* e como os nossos psiquiatras A e B, dizemos então que, $a\ b\ c\ d$ representam a proporção de casos que caem em cada uma das regiões definidas pelo limiar dos dois avaliadores. Por exemplo a é a proporção abaixo do limiar dos dois avaliadores e portanto diagnosticado como negativo pelos dois.

Estas proporções correspondem ao sumário dos dados da tabela 2x2 apresentado no quadro 5.7

Note-se que $a\ b\ c\ d$ representam proporções e não frequências.

Uma vez conhecidos os valores de $a\ b\ c\ d$ para o estudo, é simples estimar o modelo representado na 5.9.

Especificamente, estimamos a localização dos limiares de discretização, *t1, t2*, e um terceiro parâmetro *rho* que determina a largura da elipse.

Rho é o coeficiente tretracórico de correlação ou r^*.

Pode ser assim interpretado como a correlação entre a severidade da doença vista pelos avaliadores A e B antes da aplicação dos limiares.

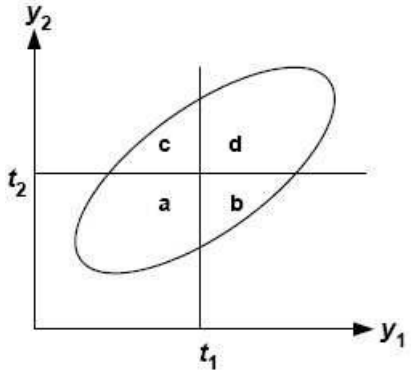


Figura 5.9: Distribuição conjunta das classificações de Y1 e Y2 e dos respectivos limiares t1 e t2

Quadro 5.7: Concordância das classificações atribuídas

		Psi A		
		Neg.	Pos.	
Psi B	Neg.	a	b	a+b
	Pos.	c	d	c+d
		a+c	b+d	

O princípio da estimação é simples: basicamente o computador tenta várias combinações para *t1, t2* e r^* até que sejam encontrados valores para os quais as proporções esperadas de *a b c d* na Figura 5.9 sejam os mais próximo possível das proporções devolvidas pelo quadro 5.7.

Os valores que assim se apresentam são considerados como estimadores dos verdadeiros parâmetros populacionais.

Na eventualidade de mais de 2 níveis na variável estima-se o coeficiente policórico, que é uma extensão do modelo ora descrito.

Conclusão

Explicado brevemente nesta secção os fundamentos da Análise Factorial, e de alguns aspectos particulares endereçados à utilização de variáveis categóricas nominais e ordinais, apresentaremos no capítulo seguinte a forma como este conceitos foram utilizados no nosso trabalho.

Na secção seguinte passaremos á apresentação, igualmente breve, do outro método a utilizar, mais propriamente a Regressão Logística.

5.2.2 Regressão Logística

Introdução

A regressão logística pertence a um grupo de modelos estatísticos denominados MLG (Modelos Lineares Generalizados).

Esta classe alargada de modelos inclui vários modelos como, a regressão linear, a ANOVA bem como modelos multi-variados como a ANCOVA, a regressão loglinear e a regressão de Poisson.

Iremos fazer aqui apenas um breve resumo do seu conceito e fundamentos, uma vez que existem vários e qualificados autores com obra publicada nesta matéria.

Destes sobressai, sem dúvida, pela qualidade e importância metodológica a obra de Hosmer e Lemeshow [34], mas também as de Agresti [2] e de Scott Long [47], entre outras.

A regressão logística permite prever um evento sob a forma de uma variável discreta, a partir de um conjunto de variáveis que podem ser contínuas (de escala, de intervalo) categóricas (binárias, ordinais) ou mistas.

Na maioria dos casos a variável resposta (habitualmente também chamada variável dependente) é uma variável dicotómica (0/1), tipo ausente/presente, insucesso/sucesso; contudo é igualmente possível construir modelos de regressão logística para variáveis multinível, como as variáveis ordinais.

O seu uso em Epidemiologia está amplamente divulgado, por um lado devido à natureza dos dados epidemiológicos, e por outro lado, devido á disponibilidade de computadores que permitem a estimação de modelos mais complexos.

O modelo

Como acima se disse a variável resposta (dependente) em regressão logística é na maioria dos casos dicotómica, isto é, pode assumir o valor 1 com uma probabilidade de sucesso de i , ou o valor 0 com uma probabilidade de sucesso de *1- i*.

Note-se, como acima se disse a eventualidade desta mesma variável poder ser ordinal, por exemplo.

Como se disse igualmente as variáveis independentes, os predictores, podem assumir qualquer forma.

Assim o modelo de regressão logística não coloca condições acerca da distribuição das variáveis independentes.

Não necessitam de ter uma distribuição normal, ter variância igual entre os grupos, etc.

A relação entre os predictores e a variável reposta não é uma função linear no modelo de regressão logística; contudo já o deverá ser em relação à transformação de *i (logit)* expresso pela equação 5.7

$$log[\Theta(x))] = log[\frac{\Theta(x)}{1-\Theta(x)}] = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \tag{5.7}$$

Esta mesma equação pode ser expressa da forma da equação 5.8

$$i = \frac{e^{\alpha+\beta_1 x_1+\beta_2 x_2+.....+\beta_k x_k}}{1+e^{\alpha+\beta_1 x_1+\beta_2 x_2+.....+\beta_k x_k}} \tag{5.8}$$

Esta equação é útil para se poder expressar o resultado da saída computacional em que as estimativas de b_k são apresentadas sob a forma de coeficientes, ou eventualmente, exponenciadas o que as converte no valor das odds-ratio.

De acordo com as recomendações de Hosmer e Lemeshow (op. citada) o modelo de regressão logística deve possuir duas características essenciais:

- ser parcimonioso, portanto contendo o mínimo de variáveis necessárias para ajustar o modelo;
- biologicamente plausível, sem o que a sua utilidade prática ficará comprometida

Para tal o modelo é criado com a inclusão de todos os predictores considerados úteis para explicar a resposta.

Ainda segundo os mesmos autores dever-se-ão estudar as interacções entre os predictores, não sendo contudo razoável considerar interacções superiores a segunda ordem, devido á sua difícil interpretação biológica [34]

Uma das técnicas utilizadas para este fim denomina-se "stepwise regression" e, consiste em construir um modelo com todas as variáveis (backward) ou só com a constante (upward) a que se vai retirando ou acrescentado automaticamente, a partir de um valor limiar previamente definido os restantes predictores.

Embora referido em todas as obras citadas não é técnica que habitualmente utilizemos.

Após construído o modelo inicial torna-se indispensável proceder ao seu ajustamento, isto é, garantir que ele inclui as variáveis explicativas necessárias e que o seu desempenho global é o melhor possível.

Para isso utilizam-se técnicas que avaliam quer o ajustamento global do modelo, quer a validade de cada uma das variáveis predictoras.

Esse será o tema da próxima secção.

Ajustamento do modelo

Ajustamento global Para efeito do ajustamento global do modelo, isto é saber saber se o modelo actual tem um melhor comportamento que um outro modelo, apresentamos três critérios habitualmente utilizados:

Os resíduos de Pearson,

O likelihood ratio test (razão de verosimilhança) e o AIC (Akaike Information Criterion)

O teste de Hosmer and Lemeshow.

Os **resíduos de Pearson**, como definidos por Hosmer and Lemeshow (op. citada) são da forma para qualquer *padrão de covariatos* da equação 5.9

$$X^2 = \sum r(y_j, \hat{\pi_i})^2 \tag{5.9}$$

O **likelihood ratio test (LR)** usa o valor maximizado da função de verosimilhança para o modelo saturado (L1,versus o mesmo valor para o modelo actual (presumivelmente mais simples).

A sua fórmula é expressa na equação 5.10

$$-2log(\frac{L_0}{L_1}) = -2[log(L_0) - log(L_1)] = -2(L_0 - L_1) \tag{5.10}$$

O resultado segue uma distribuição de χ^2 com o número de graus de liberdade igual à diferença do número de coeficientes dos dois modelos cuja interpretação é a habitual nos testes de hipóteses, considerando usualmente um α a 0.05.

Note-se um pormenor essencial que é o de os modelos necessitarem estar encadeados, isto é, as variáveis de um só poderem ser a mais ou a menos das do outro, não podem ser diferentes.

A violação deste princípio invalida o uso do método referido.

Este valor é igualmente denominado de ***Desvio*** e apresentado em saídas computacionais como **(D)**.

0 **Akaike Information Criterion (AIC)** é uma medida que permite comparar modelos que se ajustam a um determinado conjunto de dados, sem que estes estejam encadeados, bastando que a variável resposta possua a mesma distribuição.

A sua determinação tem como base a log-verosimilhança maxima que como sabemos é uma medida estatística indicadora do ajustamento global do modelo (baixos valores indiciam pior ajustamento), e p o numero de parâmetros associados ao modelo.

É dado pela fórmula 5.11:

$$AIC = -2L\left(\hat{\beta}\right) + 2\left(p+1\right) \tag{5.11}$$

0 valor do AIC é penalizado pela adição de parâmetros.

0 AIC tem como objectivo seleccionar dentre os modelos adequadamente ajustados áqueles que possuam um número mínimo de parâmetros.

Quanto menor for o valor do AIC, melhor será o modelo.

Destacam-se como vantagens do AIC, as seguintes:

- permitir de maneira objectiva determinar dentre os modelos que oferecem um ajustamento adequado qual o mais parcimonioso de fácil cálculo e interpretação
- no caso particular de MLGs, permite comparar diferentes modelos com diferentes funções de ligação, ao passo que a utilização de métodos sequenciais (backward, forward, stepwise) pode conduzir a modelos distintos com o mesmo conjunto de variáveis predictoras
- o AIC é um critério transversal que permite comparar quaisquer modelos, bastando que a variável resposta possua a mesma distribuição

Todavia não podemos considerar a abordagem pelo AIC como uma panaceia.

Apesar do AIC permitir comparar modelos distintos e ordena-los destacando o melhor dentre o grupo de modelos estudados, tal não inviabiliza a existência de um modelo melhor fora desse grupo.

O teste de Hosmer and Lemeshow A estatística de Hosmer and Lemeshow, avalia o ajustamento do modelo criando 10 grupos ordenados de indivíduos e comparando o número actual em cada grupo (observado) com o número estimado pelo modelo de regressão logística.

A estatística de teste é um qui-quadrado, interpretado como, quando não significativo, que o modelo previsto não se afasta do modelo observado.

Os 10 grupos ordenados são criados com base na sua probabilidade estimada; aqueles que têm uma probabilidade estimada abaixo de 0.1 formam um grupo e assim sucessivamente até os que apresentam probabilidades na casa dos 0.9-1.0.

Cada uma destas categorias é então dividida em dois grupos baseados no resultado observado na variável resultado (sucesso/insucesso).

As frequências esperadas para cada uma das células são obtidas a partir do modelo.

Se o modelo ajustar bem então a maioria dos indivíduos com sucesso será classificado nos decis superiores de risco e os com insucesso nos decis inferiores de risco.

Ajustamento individual

Para efeitos da avaliação da significância de cada um dos coeficientes no modelo β utilizamos o teste de Wald.

Este calcula uma estatística Z que é da forma da equação 5.12

$$Z = \frac{\hat{\beta}}{SE_{\beta}} \tag{5.12}$$

Este valor é elevado ao quadrado produzindo uma estatística de Wald com distribuição de qui-quadrado.

Esta estatística é bastante utilizada e integrada em todas as saídas computacionais, embora existam autores (como habitualmente) que apresentam algumas reservas ao seu uso [2].

Para terminar esta secção gostaríamos de apresentar uma ferramenta que iremos usar e que foi implementada por Scott Long e Jeremy Freese para ser utilizada com o STATA.

Trata-se do comando *fitstat* integrado no conjunto *SPost*, conforme explicado no texto dos autores [48].

Através do mesmo é possível comparar as estatísticas atrás citadas em dois modelos que estejamos a estudar, podendo assim decidir qual deles é o mais correcto.

Na Figura 5.10 apresentamos a saída de um dos nossos exemplos, sendo que a sua utilização será melhor explicitada mais á frente.

Tratou-se de um estudo sobre a empregabilidade das mulheres em que se avaliava se estavam ou não empregadas versus um conjunto de predictores, como o numero de filhos abaixo dos 6 anos, entre os 6 e os 18, a escolaridade do marido, a idade da mulher, o rendimento da família, etc.

O que fizemos foi um primeiro modelo de regressão só com alguns predictores, cujas estimativas gravámos sob o nome de *m1* e, depois um outro modelo já com mais preditores.

```
. fitstat, using(m1)

Measures of Fit for logistic of lfp

                           Current            Saved       Difference
Model:                    logistic         logistic
N:                             753              753                0
Log-Lik Intercept Only      -514.873         -514.873            0.000
Log-Lik Full Model          -461.133         -470.074            8.940
D                        922.267(746)     940.148(748)       17.881(2)
LR                       107.480(6)        89.599(4)         17.881(2)
Prob > LR                       0.000            0.000            0.000
McFadden's R2                   0.104            0.087            0.017
McFadden's Adj R2               0.091            0.077            0.013
ML (Cox-Snell) R2               0.133            0.112            0.021
Cragg-Uhler(Nagelkerke) R2      0.178            0.151            0.028
McKelvey & Zavoina's R2         0.188            0.155            0.033
Efron's R2                      0.137            0.115            0.022
Variance of y*                  4.050            3.891            0.158
Variance of error               3.290            3.290            0.000
Count R2                        0.673            0.663            0.011
Adj Count R2                    0.243            0.218            0.025
AIC                             1.243            1.262           -0.018
AIC*n                         936.267          950.148          -13.881
BIC                         -4019.286        -4014.653           -4.633
BIC'                           -67.735          -63.103           -4.633
BIC used by Stata              968.635          973.268           -4.633
AIC used by Stata              936.267          950.148          -13.881

Difference of    4.633 in BIC' provides positive support for current model.

Note: p-value for difference in LR is only valid if models are nested.
```

Figura 5.10: O comando fitstat (exemplo)

Corremos então o comando que apresentou os resultados da Figura 5.10 onde observamos um conjunto de estatísticas lado a lado para o modelo inicial (gravado) e para o actual.

Lá estão todas as referidas, mais alguns de que não falámos.

Note-se no fim da saída a indicação de qual o modelo a preferir baseado, neste caso, no **BIC (Bayes Information Criterion)**, embora também lá esteja o **AIC**.

O modelo que os apresenta mais pequenos é o preferido.

Note-se igualmente um ultimo alerta para o facto da necessidade do encadeamento dos modelos.

Voltaremos a estes comandos mais á frente nos resultados.

Medidas de diagnóstico

Hosmer and Lemeshow (op. citada) propõem três medidas úteis para o diagnóstico do ajustamento do modelo, particularmente para a análise gráfica dos resíduos do mesmo.

São as seguintes:

$$\Delta X^2 = \frac{r_j^2}{(1-h_j)} = r_{sj}^2 \tag{5.13}$$

Representa o decréscimo do valor da estatística de qui-quadrado de Pearson respeitante á eliminação de indivíduos com o **padrão de covariatos** $\mathbf{x}_j$.

O conceito de padrão de covariatos é importante em regressão logística e refere-se ao conjunto de indivíduos que partilham o mesmo valor de uma dada variável.

$$\Delta D_j = \frac{d_j^2}{(1-h_j)} \tag{5.14}$$

Esta estatística de influência corresponde á variação do valor da desviância motivada igualmente pela eliminação de uma *matriz de covariatos*.

$$\Delta\hat{\beta}_j = \frac{r_j^2}{(1-h_j)^2} = \frac{r_{sj}^2 h_j}{(1-h_j)} \tag{5.15}$$

Esta terceira estatística de influência examina o efeito que a eliminação dos indivíduos com uma dada *matriz de covariatos* exerce nos coeficientes estimados e nas medidas de ajustamento globais com $\mathbf{X}^2$ e $\mathbf{D}$.

5.2.3 Análise gráfica de resíduos

Os autores que estamos a citar recomendam a análise gráfica dos resíduos após a regressão.

De uma forma geral estes gráficos apresentam no eixo das abcissas os valores de $\hat{\pi}_j$ e no eixo das ordenadas qualquer uma das medidas de influência definidas na secção anterior.

Os autores recomendam a utilização de pelo menos uma delas e, outro gráfico em que no eixo das abcissas o valor da probabilidade é substituída pela estatística h_j denominada *leverage* correspondente ao elementos diagonais da matriz $\mathbf{H}$, denominada *Hat matrix*.

Valores elevados de h_j correspondem às observações que exercem influencia no modelo, notando-se que:

$$\sum h_j = p + 1 \tag{5.16}$$

ou seja o traço da matriz **H** corresponde ao número de parâmetros presentes no modelo.

No capítulo de apresentação de resultados veremos alguns deste gráficos.

5.2.4 Uma ultima nota

Uma questão complementar a ter em consideração põe-se quando um, ou mais, dos preditores são variáveis continuas.

Nesse caso necessitamos garantir que a relação do predictor é linear em toda a gama de valores da variável contínua e a sua relação constante.

Imagine-se o exemplo seguinte retirado de um trabalho por nós realizado e não publicado.

Estuda-se o evento "doença coronária" considerando um conjunto de predictores um dos quais é a idade, aqui representada por uma variável contínua.

Ora se a *odds* de contrair doença coronária não for constante em todas as idades o modelo ficará mal especificado.

Nesse caso dever-se-á considerar a idade categorizada em dois ou mais grupos e introduzi-la no modelo como uma variável discreta com o número de níveis necessário.

Existem várias formas de atingir este objectivo, mas uma bastante comum é a de "partir" a variável em vários quartis e fazer a regressão na variável resposta por cada um desses quartis.

Se as *odds* forem idênticas em todos os níveis então a variável poderá ser utilizada como variável contínua.

De outra forma deverá ser discretizada no ponto, ou pontos de corte que se consideraem mais apropriados.

Atente-se na Figura 5.11:

Os pontos marcam os valores do logaritmo da *odds* ao longo das várias idades.

Caso a relação fosse constante seguiria a linha em cheio.

Como vemos parece existirr um valor mais baixo até cerca dos 50/55 anos e depois um aumento substancial, o que inviabiliza o pressuposto de linearidade.

Neste caso deveríamos fazer um "corte" por esta idade e introduzir a variável no modelo, como atrás se disse.

O quadro 5.8 apresenta os valores que geraram o gráfico da Figura 5.11

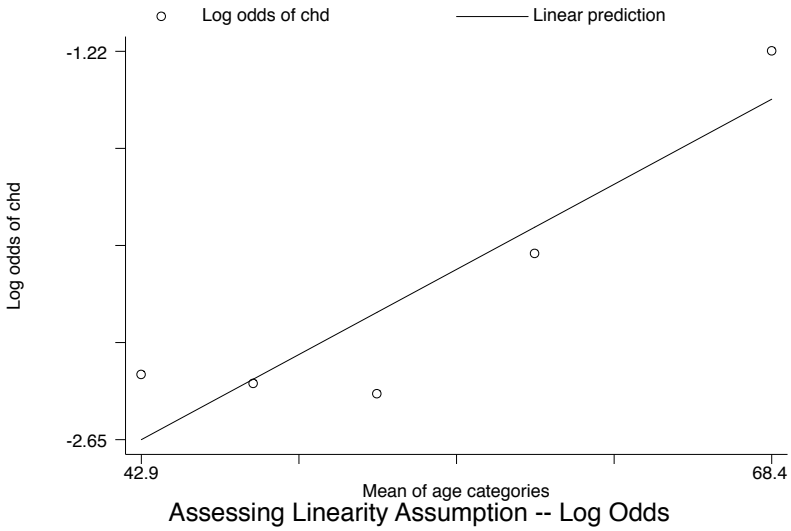


Figura 5.11: Verificando a relação linear

```
. lintrend chd age, group(5) plot(log)

The proportion and log odds of chd by categories of age

  (Note: 5 age categories of equal sample size;
     Uses mean age value for each category)
```

age	min	max	d	total	chd	logodds
42.9	40	45	11	134	0.08	-2.41
47.4	46	49	9	113	0.08	-2.45
52.4	50	55	10	130	0.08	-2.48
58.8	56	63	14	114	0.12	-1.97
68.4	64	76	27	118	0.23	-1.22

Quadro 5.8: Os dados da figura

Note-se que entre os 50 e os 55 existe uma variação da *log-odds* de cerca de -2.40 para cerca de -1.97, o que corresponde ao aumento verificado na Figura 5.11.

Estes resultados foram obtidos através de mais um programa escrito para o STATA por uma investigadora da Universidade de Washington de seu nome Joanne Garret.

É feito justamente utilizando os conceitos acima definidos e divide a variável contínua em vários grupos à escolha do utilizador, fazendo depois a respectiva regressão.

Neste caso foram utilizados 5 grupos conforme se pode observar no código inserido no quadro 5.8.

Estão referidas outras formas de contornar este problema, como a introdução no modelo de termos transformados por potenciação, radiciação, etc ou a utilização de Modelos Generalizados Aditivos (GAM), técnicas mais sofisticadas que estão fora do nosso âmbito.

5.3 Marcha Geral da Análise

No nosso caso, tivemos de tomar um conjunto múltiplo de decisões para conseguir fazer o estudo que necessitam de uma cabal e pormenorizada explicação.

Assim, para evitar sobrecarregar o capítulo de resultados com a discussão destas decisões, optou-se pela sua explanação aqui, salvo algum pormenor que seja preciso considerar mais á frente.

Desta forma consideramos, a partir deste capítulo, explicadas as opções tomadas e as razões que a isso levaram.

Essas opções foram várias, desde a escolha dos métodos, a forma de trabalhar os dados, até á escolha do software utilizado.

Começaremos por esta última opção.

5.3.1 *Software* utilizado

Os dados para o trabalho chegaram ás nossas mãos sob a forma de folha de cálculo formato *.csv* com campos delimitados por ponto e vírgula, com as características que salientámos no Capítulo 5 e que referiremos na secção seguinte mais em pormenor.

Tornava-se necessário transferi-los para um programa de estatística a fim de os poder processar.

Três opções se nos ofereciam, nomeadamente SPSS, STATA e R.

A nossa escolha recaiu no STATA porque este tem a possibilidade de fazer Análise Factorial a partir de matrizes de correlação tetracórica, quer com o comando base do STATA, quer recorrendo a programas escritos por utilizadores.

No caso presente utilizámos o programa *polychoric* escrito por Stas Kolenikov e Gustavo Angeles conforme descrito no artigo original [43].

De acordo com comentários de vários utilizadores, os coeficientes obtidos com este programa são mais fiáveis do que os mesmos obtidos com o comando base, razão para a sua utilização.

Ainda de referir que no domínio deste programas escritos pelos utilizadores, ressaltam os programas de Scott Long e Jeremy Freeze, bem como o de Joanne Garrett descritos sumariamente no Capítulo 5, e que iremos utilizar algumas vezes.

5.3.2 Análise de dados

Conforme dissemos na secção anterior os nossos dados foram recebidos em formato de folha de cálculo *.csv* com delimitador de ponto e vírgula.

A leitura do quadro 5.1 e seguintes mostram-nos as variáveis do GDH; as que mais nos interessam *(ddx1, ddx20)* contém os diagnósticos dos vários episódios codificados de acordo com a CID-9.

Recordamos igualmente que o nosso objectivo é estudar a multimorbilidade em doentes com internamento hospitalar entre um dia e um ano, pelo que duas tarefas iniciais se levantam.

Primeiro eliminar os episódios com menos de um dia e mais de um ano; de seguida converter as variáveis dos diagnósticos, que se apresentam em formato alfanumérico, em variáveis com formato utilizável pelo programa, neste caso variáveis binárias codificadas 0/1, conforme o diagnóstico estiver presente ou ausente.

Para tal, e porque com o Excel tivemos dificuldades em manusear tão grande folha de dados, convertemos ano a ano, cada folha para o formato *.sav* do SPSS através do filtro de importação do programa.

Uma vez importados os dados foi então fácil através dos comandos habituais do SPSS eliminar as observações que não nos interessavam.

Ficamos assim com três folhas de dados com respectivamente 307881 306256 e 320581 observações respectivamente em 2009, 2010 e 2011.

O próximo passo consistiu então em converter os diagnósticos em variáveis binárias 0/1; para tanto foi escrita uma rotina em SPSS (pelo Prof. Paulo Nogueira, uma vez que entre as nossas várias limitações existe a de não sermos programador informático) que "varreu" toda a base de dados atribuindo a cada uma das *ddx* o valor 0/1 conforme se observava um dos 999

diagnósticos da CID-9, criando assim um conjunto de variáveis denominada de *var_ 001, var_ 999* que passaram a ser o nosso material de trabalho.

O quadro 5.9 apresenta uma pequeníssima parte dessa rotina.

Como vemos esta retira de cada *ddx* 3 caracteres a contar do último com o comando *CHAR.SUBSTR(ddx...1,3)* e coloca-o numa variável auxiliar com as características de variável numérica (topo da rotina).

Este processo é repetido 999 vezes a fim de correr todos os diagnósticos da CID-9.

Foram excluídos deste processo o 99 diagnósticos correspondentes á parte final da CID-9 e que dizem respeito a procedimentos e outros aspectos exteriores, uma vez que este diagnósticos representam, na sua grande maioria, os casos com menos de um dia de internamento que tínhamos eliminado no passo anterior.

Uma vez terminada esta operação exportou-se a base de dados para o STATA versão 12 manendo-se as seguintes variáveis:

```
STRING aux (A3).
COMPUTE aux="001".
NUMERIC  var_001 (F1.0).
COMPUTE var_001 = CHAR.SUBSTR(ddx1,1,3)=auxlCHAR.SUBSTR(ddx2,1,3)=auxl
CHAR.SUBSTR(ddx3,1,3)=auxlCHAR.SUBSTR(ddx4,1,3)=auxlCHAR.SUBSTR(ddx5,1,3)=auxl
CHAR.SUBSTR(ddx6,1,3)=auxlCHAR.SUBSTR(ddx7,1,3)=auxl
CHAR.SUBSTR(ddx8,1,3)=auxlCHAR.SUBSTR(ddx9,1,3)=auxlCHAR.SUBSTR(ddx10,1,3)=auxl
CHAR.SUBSTR(ddx11,1,3)=auxlCHAR.SUBSTR(ddx12,1,3)=auxl
CHAR.SUBSTR(ddx13,1,3)=auxlCHAR.SUBSTR(ddx14,1,3)=auxl
CHAR.SUBSTR(ddx15,1,3)=auxlCHAR.SUBSTR(ddx16,1,3)=auxl
CHAR.SUBSTR(ddx17,1,3)=auxlCHAR.SUBSTR(ddx18,1,3)=auxl
CHAR.SUBSTR(ddx19,1,3)=auxlCHAR.SUBSTR(ddx20,1,3)=aux.
EXECUTE.
```

Quadro 5.9: Sintaxe do SPSS

- Ano (ano)
- Sexo (sexo)
- Idade (idade)
- Dias de Internamento (dias_int)
- Tipo de admissão (adm_tip)
- Destino após Alta (dsp)

- var_001-var_999

Entendeu-se que as restantes variáveis constantes do GDH não se revelavam portadoras de informação relevante para o nosso estudo.

As variáveis *ano, adm_tip, dsp* são variáveis nominais, sendo que as variáveis *idade, dias_int* são contínuas e as *var_001 a var_999* são binárias dicotómicas.

O passo seguinte revelou-se particularmente problemático e de decisão difícil.

Tentando uma primeira análise dos dados esta revelou-se confrangedoramente incoerente, sem que se conseguisse visualizar qualquer padrão minimamente aceitável.

Rapidamente percebemos que estávamos perante um problema de grande diferença no número de observações nas diferentes variáveis.

A razão prende-se com dois factos, no nosso ponto de vista, a saber:

1. A existência de códigos de diagnóstico muito semelhantes, comportando-se como variações de um diagnóstico principal;

2. A existência de codificação muito atomizada, o que deverá prender-se com a necessidade de ser minucioso na descrição dos actos praticados para efeito da devida remuneração

Desta forma pareceu sensato remover as variáveis que apresentassem um número reduzido de observações.

Experimentámos subtrair variáveis com menos de 100, 500 e 1000 observações.

O impacto na base de dados pareceu favorável á manutenção de variáveis com mais de 1000 observações, opção que assumimos desde então.

Foi então realizada a Análise Factorial para cada um dos anos, com base na matriz de correlação tetracórica como acima se disse.

Os factores foram retidos de acordo com a regra de Kaiser e a análise do *screeplot* como atrás explicado.

Como parece razoável admitir que as várias patologias se encontram ligadas entre si, optou-se por realizar uma rotação oblíqua (*oblimin*); considerámos então as variáveis que apresentavam *loadings* superiores a .25 como no trabalho de Schaffer (op. citada).

Os nossos dados foram um pouco menos consistentes do que os desse autor, uma vez que ele lida com observações individuais, e nós com episódios de urgência que podem ocorrer em doentes isolados ou no mesmo doente. Verificámos nesta fase um facto que reputamos de importante. Um conjunto de variáveis relacionadas com a gravidez e o parto encontram-se

consistentemente associadas constituindo assim um *cluster* muito homogéneo.

Ora este conjunto de variáveis está um pouco para além do nosso estudo, pelo que nos pareceu ajuizado não as considerar e portanto excluir os grupos 740-759 e 630-679 da CID-9.

Admitimos que possa existir comorbilidade nesta população mas, a existir, será um grupo especial que não parece avisado juntar com a população geral.

Identificadas as variáveis construímos então variáveis compostas correspondentes a cada um dos factores, da seguinte forma.

Se chamarmos a cada factor F_x e aos *loadings* das respectivas variáveis l_{vark} cada factor ficará então definido da forma:

$$F_x = l_{var1} + l_{var2} + ... + l_{vark} \tag{5.17}$$

relembrando que l_{vark} é maior que 0.25.

Construímos de seguida variáveis dicotómicas que representem as situações em que estiverem presentes 2 ou mais diagnósticos.

O código é o seguinte:

$$recodeF1(min/1 = 0)(2/max = 1), gen(p1) \tag{5.18}$$

Repetimos esta operação para todos os factores gerando assim as variáveis $p_1....pk$.

Finalmente como pretendemos considerar multimorbilidade qualquer combinação de dois ou mais diagnósticos em qualquer dos factores construímos uma variável que reflecte esse facto da forma:

$$genp_{multi} = p_1|p_2|p_3|....p_k \tag{5.19}$$

Assim a variável p_{multi} representa essa combinação e será integrada no modelo de regressão logística como variável de resposta.

Este modelo será construído, inicialmente, com todas as variáveis independentes, considerando:

- O teste de linearidade das variáveis contínuas no *logit* ao longo de todos os seus valores, usando o programa *lintrend*.
- Todas as interacções de primeira ordem entre elas;
- Todos os coeficientes com *p-value* supeiror a 0.25 pelas razões explicitadas na obra de Hosmer e Lemeshow [34].

O ajustamento do modelo será avaliado pelo **Critério de Informação de Akaike (AIC)** até se obter o melhor resultado (valor mais baixo).

Usaremos o programa de Scott Long já citado [48].

Procederemos igualmente á análise gráfica dos resíduos como atrás se disse.

Capítulo 6

Resultados

Vamos apresentar de seguida os resultados; refira-se que estes dizem respeito a episódios de internamento.

Para 2011 foi possível ir um pouco mais além, uma vez que o GDH passou a incluir um *número fictício de utente gdh* que nos permitiu agregar as observações pelos respectivos indivíduos.

Assim para esse ano iremos apresentar os resultados das duas formas, utilizando esta ultima (dados agregados) igualmente para construir o modelo de regressão logística.

Apresentamos quadros resumo comparativos dos três anos.

Em anexo encontram-se as respectivas saídas do computador completas.

6.1 Abordagem Descritiva

No ano de 2009 registámos 307881 episódios de internamento nas condições previamente definidas.

No ano de 2010 contabilizámos 306318 episódios e no de 2011 contabilizámos 231809. Estes referem-se a dados agrupados da forma acima referida.

Registámos, para além das variáveis com diagnósticos em que se contabilizaram mais de mil (1000) observações, as variáveis, aliás comuns a todos os anos que apresentamos no quadro 6.1

Vamos proceder a uma análise univariada das variáveis contínuas e das variáveis nominais e de seguida uma análise bivariada entre elas.

6.1.1 Análise univariada

Idade e dias de Internamento Analisámos então, isoladamente as variáveis contínuas cujo resultado se apresenta no quadro 6.2

Quadro 6.1: Codificação das variáveis

Variáveis				
idade	sexo	dias_int	dsp	adm_tip
	1=M		0=Desconhecido	1=Programado
	2=F		1=Domicílio	2=Urgente
	3=Ind.		2=Outro Hosp.	3=Acesso
			6_Serv. Dom.	4=PECLEC
			7=Cntr. Parecer	5=Privada
			20=Flecido	6=SIGIC
				7=PACO

O tipo de admissão foi assim recodificado (1 3 4 6)=Programado, 20=urgente, 5=Privado, 7=PACO

Vemos que a média da idade é de 48.43 anos com um desvio-padrão de 27.9 em 2009, 48.64 anos com um desvio-padrão de 28.29 em 2010 e de 50.82 anos com um desvio-padrão de 26.16 em 2011.

Quanto aos dias de internamento a sua média é de 7.76 com um desvio-padrão de 12.27 em 2009, 7.78 com um desvio-padrão de 12.23 em 2010 e de 7.85 com um desvio-padrão de 11.86 em 2011.

Esta variável apresenta um desvio á direita com a presença de múltiplos *outliers* em todos os anos, como se pode exemplificar pelo gráfico de dispersão da Figura 6.1, referente ao ano de 2009.

Idade			
	2009	2010	2011
Media	48.43	48.64	50.82
D. P.	27.99	28.29	26.16
Dias de Internamento			
	2009	2010	2011
Media	7.76	7.78	7.85
D.P:	12.27	12.23	11.86

Quadro 6.2: Idade e Dias de Internamento

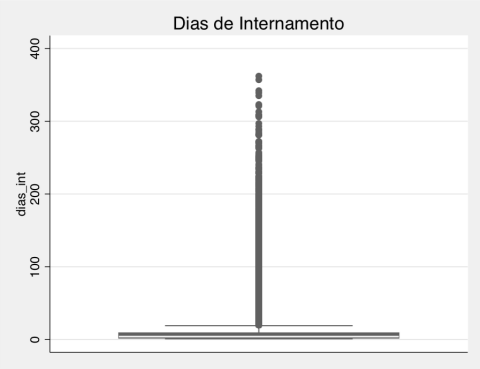


Figura 6.1: Gráfico dos dias de internamento

Sexo, destino após alta e tipo de admissão Quanto ao sexo, e em relação a 2009, verificamos que 55.29% são mulheres enquanto 44.71% são homens.

No ano de 2010 verificamos que 55.36% são mulheres enquanto 44.64% são homens.

No ano de 2011 57.64% são mulheres enquanto 42.57% são homens

O destino após alta em 2009 é preferencialmente o domicilio 92% ficando os restantes 8% para os restantes destinos.

Em 2010 o destino após alta é preferencialmente o domicilio 92% ficando os restantes 8% para os restantes destinos, em em 2011 o destino após alta é igualmente preferencialmente o domicilio 92% ficando os restantes 8% para os restantes destinos.

O tipo de admissão 66.52% é urgente e 33.4% programado em 2009, seguindo um perfil semelhante em 2010 e 2011 com 67,9% urgente e 32,0% programado e 66.02% urgente e 33.98% programado, respectivamente

Os resultados estão no quadro 6.3 e quadro 6.4

6.1.2 Análise bivariada

Fomos agora verificar se a idade se distribuía igualmente pelos vários sexos e, se os dias de internamento se distribuíam igualmente nos dois sexos.

Lembramos que, muito embora a variável sexo esteja codificado em três níveis; como o quadro 6.1 mostra o número de observações com sexo indeterminado é tão baixa que nos pareceu razoável descarta-las, até porque não temos conhecimento das razões que levaram a essa classificação.

Quadro 6.3: Sexo e tipo de admissão

	Sexo					
	2009		2010		2011	
	Freq.	%	Freq.	%	Freq.	%
Masc.	137635	44.71	136698	44.64	98970	42.54
Fem.	170214	55.29	169520	55.36	133702	57.46

	Tipo de Admissão					
	2009		2010		2011	
	Freq.	%	Freq.	%	Freq.	%
Program.	103052	33.48	98023	32.01	79064	33.98
Urgente	204794	66.52	208195	67.99	153606	66.02

Quadro 6.4: Destino após alta

	Destino após alta					
	2009		2010		2011	
	Freq.	%	Freq.	%	Freq.	%
Dom.	283221	92.0	281104	91.80	215301	92.53
Outro Hosp.	7765	2.52	7795	2.55	4950	2.13
Serv. Dom.	385	0.13	1003	0.33	468	0.20
Cntr Parecer	1714	0.56	1425	0.47	1113	0.48
Falecido	14738	4.79	14891	4.86	10.447	4.49

O resultado para a idade e dias de internamento por sexo está presente no quadro 6.5

Em relação a 2009 vemos que existem diferenças nos dois sexos; os homens são mais idosos: média 49.52, IC(49.37-49.67) contra 47.54 IC(47.41-47.67) anos nas mulheres.

Os tempos de internamento são também maiores: homens média 8.51 IC(8.43-8.58) contra 7.16 IC(7.11-7.21) nas mulheres.

Em relação a 2010 vemos que existem diferenças nos dois sexos; os homens são mais idosos: média 49.76, IC(49.61-49.92) contra 47.73 IC(47.60-47.86) anos nas mulheres.

Quadro 6.5: Idade e Dias de Internamento por sexo

	Idade por Sexo					
	2009		2010		2011	
	Media	$\Pr(\|T\| > \|t\|)$	Media	$\Pr(\|T\| > \|t\|)$	Media	$\Pr(\|T\| > \|t\|)$
Masc.	49.52	0.000	49.76	0.000	52.37	0.000
Fem.	47.54		47.73		49.60	

	Dias de Internamento por Sexo					
	2009		2010		2011	
	Media	$\Pr(\|T\| > \|t\|)$		$\Pr(\|T\| > \|t\|)$		$\Pr(\|T\| > \|t\|)$
Masc.	8.51	0.000	8.51	0.000	8.43	0.000
Fem.	7.16		7.12		6.79	

Os tempos de internamento são também maiores: homens média 8.519 IC(8.52-8.66) contra 7.12 IC(7.06-7.17) nas mulheres.

Para 2011 existem diferenças nos dois sexos; os homens são mais idosos: média 52.37 IC(52.21-52.54) contra 49.60 IC(49.46-49.74) anos nas mulheres.

Os tempos de internamento são também maiores: homens média 8.43IC (8.35-8.51) contra 6.79 IC(6.74-6.85) nas mulheres.

O teste *t de Studen*t efectuado mostra que estas diferenças são estatisticamente significativas, como se pode igualmente observar no quadro 6.5

Em relação á diferença de idade no que se refere ao destino após alta e aos dias de internamento por destino após alta verificamos igualmente a existência de diferenças como se pode observar no quadro 6.6

Podemos observar que a saída para outras instituições, serviços domiciliários e óbitos estão associados a idades mais avançadas; a saída contra parecer está porém associada a idades mais jovens, o que parece vir de encontro á constatação feita na prática clínica diária.

O destino após alta também está relacionado com os dias de internamento; o serviço domiciliário é neste caso o que apresenta maior tempo de internamento, seguido dos óbitos e de outras instituições.

Os resultados são consistentes nos três anos.

Refira-se igualmente que no ano de 2011 foram introduzidos novos destinos após alta, conforme referido na tabela 6.1; simultaneamente o destino

Quadro 6.6: Idade e Dias de Internamento por Destino após Alta

Idade por Destino após Alta			
	2009	2010	2011
Domicilio	46.91	47.07	49.50
Outro Hosp.	56.00	56.00	56.72
Serv. Dom.	43.68	59.18	–
Cntr Parecer	43.65	41.84	43.48
Falecido	73.75	74.32	75.16

Dias Internamento por Destino após Alta			
	2009	2010	2011
Domicilio	7.21	7.21	6.95
Outro Hosp.	11.08	10.78	12.45
Serv. Dom	37.85	25.44	–
Cntr Parecer	6.78	6.42	6.44
Falecido	15.93	15.89	16.37

Serviço Domiciliário foi suprimido. Estes novos destinos, porém, apresentavam uma número de casos muito reduzidos pelo que resolvemos eliminá-los, até para não levantar problemas com posteriores testes de comparação.

Corremos uma *oneway ANOVA* seguida de *teste de comparação de Sidak* que nos permitiu concluir que as diferenças encontradas são estatisticamente significativas.

Quanto ao tipo de admissão os episódios de urgência acontecem em idades mais jovens, o que neste caso contraria um pouco a experiência clínica acima referida; contudo devemos recordar que estamos a lidar com episódios de urgência, não com indivíduos, o que pode introduzir algum erro nesta avaliação.

Quanto aos dias de internamento por tipo de admissão verifica-se igualmente que os episódios provenientes da urgência têm maior tempo de internamento.

Igualmente o teste *t de Student* utilizado demonstrou ser essa diferença estatisticamente significativa.

Os resultados estão no quadro 6.7

Igualmente fomos analisar o sexo versus o destino após alta e versus o tipo de admissão.

Quadro 6.7: Idade e Dias de Internamento por Tipo de Admissão

Idade por Tipo de Admissão						
	2009		2010		2011	
	Media	Pr($\|T\| > \|t\|$)	Media	Pr($\|T\| > \|t\|$)	Media	Pr($\|T\| > \|t\|$)
Programada	52.11	0.000	51.75	0.000	51.17	0.000
Urgente	46.59		47.17		50.58	
Dias de Internamento por Tipo de Admissão						
	2009		2010		2011	
	Media	Pr($\|T\| > \|t\|$)	Media	Pr($\|T\| > \|t\|$)	Media	Pr($\|T\| > \|t\|$)
Programada	5.67	0.000	5.40	0.000	5.09	0.000
Urgente	8.87		8.89		8.73	

Os resultados estão no quadro 6.8 e quadro 6.9

Quadro 6.8: Sexo e Destino após Alta

Sexo e Destino Após Alta				
2009	Dom.	Outro Hosp.	Cntr. Parecer	Falecido
Masc.	123961	4773	816	7896
Fem.	159250	2992	898	6842
2010				
Masc.	122634	4824	671	8122
Fem.	158470	2971	754	6769
2011				
Masc.	89451	3090	595	5539
Fem.	125848	1860	608	4908

De notar que existiam cerca de 22 casos admitidos fora do contexto da urgência e programada.

Sucede que esses casos se distribuíam por quatro células, uma das quais se encontrava a zero; desta forma a validade do teste de qui-quadrado ficava

Quadro 6.9: Sexo e Tipo de Admissão

	Sexo por Tipo de Admissão					
	2009		2010		2011	
	Prg	Urg	Prg.	Urg.	Prg.	Urg.
Masc.	48939	88696	46775	89923	36187	62398
Fem.	54113	116098	51248	118272	42737	90478

comprometida pelo que decidimos suprimir essas observações, o que nos pareceu não problemático.

Podemos observar que existem diferenças estatisticamente significativas, sendo que existem mais altas para o domicilio do sexo feminino e menos óbitos deste sexo.

Quanto ao tipo de admissão as diferenças são igualmente significativas, sendo que existem mais admissões do sexo feminino quer na urgência, quer nas admissões programadas.

De uma forma geral, talvez com excepção á idade dos episódios de urgência, estes resultados combinam, como acima se disse, razoavelmente com a impressão que se obtém da prática clínica diária: mais episódios em homens, mais idosos, mais dias de internamento, mais dependentes, e com mortalidade acrescida.

Note-se que, conforme anteriormente apontado os resultados de 2011 se referem aos episódios agregados pelo *numero fictício de utente gdh*, o que justifica as diferenças encontradas em alguns quadros; no caso do sexo e tipo de admissão, por exemplo, verifica-se uma diferença evidente entre 2011 e os restantes anos.

No primeiro caso estamo-nos a referir episódios á volta de indivíduos, enquanto nos outros nos estamos a referir a episódios na sua totalidade.

6.2 Abordagem por Análise Factorial

Conforme acima foi explicado realizámos uma análise factorial a uma matriz de correlação tetracórica construída com o pacote *polychoric* do STATA.

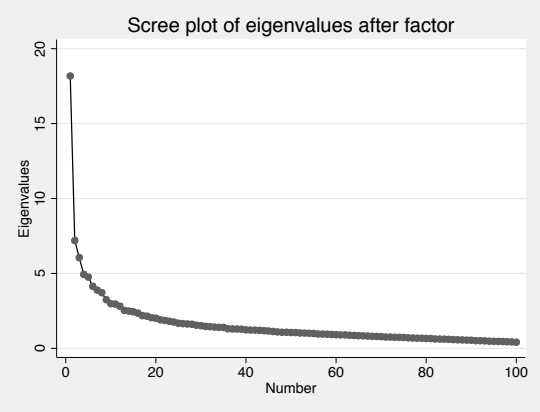
Embora todas as saídas computacionais se encontrem em anexo decidimos, para não sobrecarregar o texto principal apresentar somente os *eigenvalues* dos factores relevantes, o *scree plot* e os respectivos *loadings* nas variáveis de interesse, após rotação *oblimin*.

Os quadros 6.2 6.3 e 6.4 apresentam os *eigenvalues* e os *screeplo*t para cada um dos anos de 2009 a 2011, considerando a análise dos episódios.

No quadro 6.5 apresentamos os mesmos resultados para 2011 considerando os dados agregados.

Quadro 6.2: Os *eigenvalues* de 2009

Factor	*Eigenvalues*	Proporção
Factor1	18.17	0.10
Factor2	7.19	0.04
Factor3	6.05	0.03
Factor4	4.93	0.02
Factor5	4.75	0.02
Factor6	4.13	0.02
Factor7	3.88	0.02
Factor8	3.70	0.02
Factor9	3.24	0.01
Factor10	2.97	0.01
Cumulativa		0.35

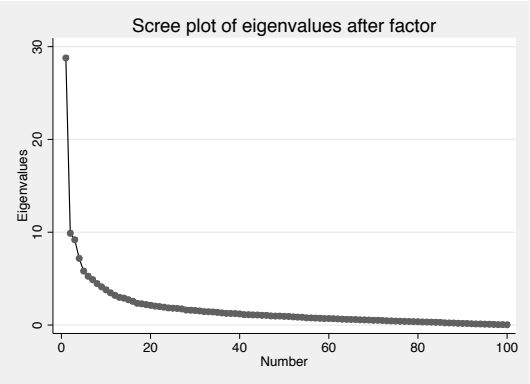


Da análise dos elementos referidos verificamos que existe um número elevado de factores com *eigenvalues* próximos de 1, o que parece habitual em dados com as características dos nossos.

Se considerássemos somente a regra de Kaiser (*eigenvalues* superiores a 1) teríamos cerca de 55 factores a maior parte dos quais com contributos

Quadro 6.3: Os *eigenvalues* de 2010

Factor	*Eigenvalues*	Proporção
Factor1	28.77	0.16
Factor2	9.88	0.22
Factor3	9.18	0.27
Factor4	7.17	0.31
Factor5	5.80	0.35
Factor6	5.25	0.38
Factor7	4.89	0.41
Factor8	4.46	0.43
Factor9	4.11	0.45
Factor10	3.79	0.48
Cmulativa		0.50



muito reduzidos para o modelo.

Nestas condições teríamos cerca de 77% de variabilidade explicada o que é manifestamente pouco para tanto factor.

Compulsando agora estes dados com os *screeplots* do quadro 6.2, quadro 6.3 e quadro 6.4 estas parecem indicar uma inflexão clara por volta dos 8/9 factores estando os restantes num patamar muito próximo.

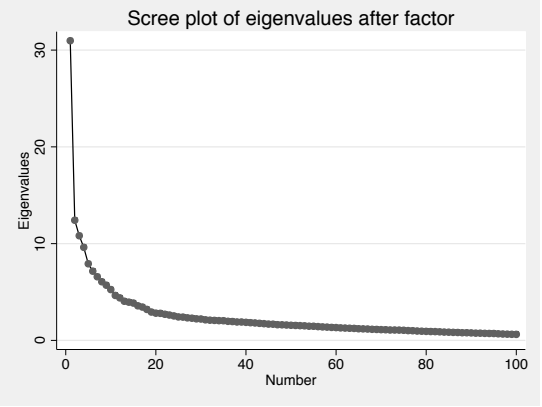
Desta forma pareceu-nos razoável reter somente os 8 primeiros factores, embora tenhamos a consciência de que estes só explicam cerca de 33% dos dados.

O caso do ano de 2011, no que respeita aos dados agregados é diferente pelo que apresentaremos um quadro a ele dedicado.

Contudo não podemos esquecer que estamos a lidar com registos de episódios, e aí talvez uma possível razão para a baixa aderência do modelo.

Quadro 6.4: Os *eigenvalue*s de 2011

Factor	*Eigenvalues*	Proporção
Factor1	30.96	0.11
Factor2	12.41	0.16
Factor3	10.81	0.20
Factor4	9.61	0.23
Factor5	7.91	0.26
Factor6	7.15	0.29
Factor7	6.57	0.32
Factor8	6.05	0.34
Factor9	5.70	0.36
Factor10	5.24	0.38
Cumultiva		0.38



Após a rotação *oblimin* destes factores obtivemos os quadros de *loadings* que se encontra no anexo I; deles extraímos os quadros resumo quadro 6.10, quadro 6.11 e quadro somente com as variáveis por grupos diagnósticos e os respectivos factores com *loadings* superiores a 0.25.

Apresentados os quadros para os três anos, mostramos no quadro 6.13 os resultados para o ano de 2011 considerando os episódios agrupados por indivíduo.

Deve notar-se que aqui, uma vez que vários episódios podem partilhar o mesmo indivíduo, as variáveis já não se apresentam como categóricas binárias mas sim de uma forma que podemos no limite considerar como contínua.

Quadro 6.10: As variáveis e os factores 2009

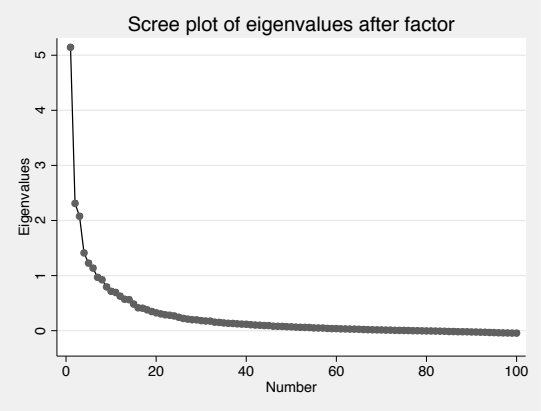
	2009
F1	Neoplasia da próstata (185);dist. hidroele. anemias (276, 280, 2819); Demência, Doença de Parkinson, epilepsia e outros distúr. cérebro (290-294, 331-332, 345-348 e 407); Doença hipertens. coração (402); bronco e pneumo(485-486); Doença renal vias urinárias (593-599 e 798)
F2	D. sangue (284-288); Lesões ouvido inte. e o sínd. vertig.(381-389);D. coração(410-417); complic. act. med(995-998)
F3	Diabetes (250); obesidade (278); alt. met. lípidos (272);HTA(401-403); D. cornar. EAM e ICC(410-428)
F4	D. figado e complic. (530-578)
F5	Ap.gen. fem. except. grav. (617-625)
F6	Neo prós. cólon, recto, ,estômago, pulmão (151-154, 162, 196-198)
F7	Dep. álcool, droga (304-305); d. pulm. crónica (415-416) e complic. d. pulm.(485-518)
F8	Depressão (311);D. Parkinson epilepsia, hemiparesia e hemiplegia (342-345); D.cerebrovasc. incl. lesões hemorrag. intracran.(431-435)

Quadro 6.11: As variáveis e os factores de 2010

	2010
F1	D. sangue(280-288);abuso de álcool e drogas (304-305) alt. fígado, inc. cirrose(571-578) osteoartrose (705-789);
F2	Asma, bronquite crónica, enfisema, (466-494) (511-519) (783-799) inc. cancro pulmão (162) efeitos tard. TP;negativo patologias do sexo feminino (618-625)
F3	Parkinson, demência (331-345); d. isque. cerebrovasc. (434-438);Insuf. card. (428), na Insufic. renal (584-599), D. osteoartic. (722-728), fract. colo fémur (820) úlcer. crónic. pele
F4	EAM e D. coronária (410-418); D. cerebrovas. aguda (785-790) Insufic. renal aguda e crónica (785-790);
F5	Diabetes alt. lípidos obesidade (250/278);HTA; d. coronar ICC (401-428); D. cerebrovasc. (433-438) Aterosclerose (440) HBP (600);
F6	D. Ap. digest., excl. o fígado; D. reuma. (714-733) (530-564)
F7	Neo estômago, cólon recto e sigmóide, pulmão (151-162) e (560-569);
F8	AVC isquémico, hemorragia e alt. várias do cérebro (431-438), consumos abusivos (303-305) alterações demenciais orgânicas (331-348)

Quadro 6.5: Os dados agregados de 2011

Factor	*Eigenvalues*	Proporção
Factor1	5.13	0.33
Factor2	2.07	0.13
Factor3	1.41	0.09
Factor4	1.22	0.07
Factor5	1.13	0.07
Factor6	0.96	0.06
Factor7	0.92	0.06
Cumulativa		0.83



O valor da medida de Kaiser-Meyer-Olkin é de 0.7941.

A partir destes resultados foram então construídas as variáveis compostas considerando a existência de 2 ou mais diagnósticos em qualquer dos factores; assim como atrás se explicou foi analisado o fenómeno da multimorbilidade cujos resultados apresentamos no quadro 6.14 .

Verificamos a existência de multimorbilidade, como a definimos, em cerca de 35% dos casos, com pequena variação anual sendo mais elevada no ano de 2009.

6.3 Modelação estatística

O modelo será constituído pelas seguintes variáveis: como variável dependente *pmulti*, como variáveis independentes *idade, sexo, dias de interna-*

Quadro 6.12: As variáveis e os factores de 2011

	2011
F1	Diabetes (250) dislipidemia (272) obesidade (278);demências (293-294),D. de Parkinson, hemiparesia, epilepsia otrs doenças cérebro (332-348);vertigens (386) def visual. inc.cataratas (368-370);hemorragia cerebral (431-438)
F2	Doença isquémica do coração (410-429);pneumonias, pleurisias, (481-519)
F3	D. estômago, pâncreas intestino e fígado (excluindo hepatite viral) (531-578)
F4	Doenças osteoarticulares (710-735)
F5	Acid. de proc. clínicos e instr. (996-998); erisipela (035) a septicemia (038) e algumas patologias do aparelho urinário, em especial das vias urinárias (599)
F6	Negativo para fracturas (802-873)
F7	Neoplasias, nomeadamente pulmão, colon, estômago, mama, etc. (150-199)
F8	Esclerose múltipla (334-340) e também nalgumas perturbações neuromusculares (729-738)

mento, destino após alta, tipo de admissão.

Note-se que temos duas variáveis contínuas *idade, dias de internamento* pelo que temos de nos assegurar da linearidade da relação com o *logit* ao longo de todos os valores da variável.

Vamos utilizar o método desenvolvido no programa *lintrend* para o STATA, como explicado no capítulo dos métodos.

Apresentamos as tabelas para a idade e para os dias de internamento na figura6.6.

Da análise da figura podemos verificar que a *logodds* não se mantém constante ao longo de todos os valores da idade.

Assim a partir dos 70 anos existe uma variação do valor da mesma que

Quadro 6.13: As variáveis e os factores de 2011 com os dados agregados

	2011 (Dados agregados)
F1	Dependência do álcool (303) com alterações hepáticas incluindo a cirrose (567-572) e distúrbios não especificados do abdómen e pélvis (789)
F2	D. renal com anemia não especificada (285) doença renal hipertensiva (403) doença cardíaca com componente renal (404) insuficiência cardíaca (428) e doença renal crónica (585)
F3	Diabetes (250), dislipidemia (272), hipertensão (401), enfarte agudo do miocárdio (410), sequelas de enfarte (412), doença isquémica do coração (414), disritmias cardíacas (427), e insuficiência cardíaca (428)
F4	Oclusão das artérias cerebrais (434), outras doenças cerebrovasculares (437), efeitos tardios da doença cerebrovascular (438), mas também á hipertensão (401), á demência (290) e a doenças da uretra (599) e ulceras de pele (707)
F5	Osteoartrose (715), lumbago (724) e sintomas envolvendo o aparelho musculoesquelético (781)
F6	Pneumonia (486), a doença pulmonar obstrutiva crónica (491), as bronquiectasias (494), o cor pulmonale crónico (416), mas também as disritmias (427), a insuficiência cardíaca (428), outras doenças do pulmão (518) e efeitos tardios da tuberculose (137)
F7	Neoplasias secundárias (198), doenças da série branca (288) e sintomas gerais (780)

Quadro 6.14: A multimorbilidade

	2009		2010		2011		2011 (agr)	
	Freq.	%	Freq.	%	Freq.	%	Freq.	%
Sem MM	193591	62.89	185084	60.44	219809	68.57	157635	68.0
Com MM	114255	37.11	121134	39.56	100766	31.43	74174	32.0
Total	307846		306218		320575		231809	

corresponde ao aumento da possibilidade de apresentar mutlimorbilidade.

Para os dias de internamento também verificamos uma modificação na relação entre os dias de internamento por volta dos 10 dias como podemos confirmar na figura 6.6

Desta feita faz sentido construir uma variável dummy que separe os dois níveis da variável, antes e depois dos 70 anos e os dias de internamento antes e depois dos 10 dias.

Procedemos á etiquetagem das novas variáveis, a fim de tornar a sua leitura mais fácil, e passámos então á construção do modelo.

Obteve-se um modelo de feitos principais inicialmente como demonstra a figura 6.7

Todas as variáveis, incluindo as variáveis construídas acima se mostraram estatisticamente significativas.

As *odds* são também elevadas, nomeadamente as da idade depois de categorizada (6.45) e os dias de internamento divididos acima de 10 dias (2.17).

Interessa agora analisar as interacções e possíveis sinais da existência de confundimento.

Para esse efeito estudaram-se então vários modelos com todas as interacções entre os preditores .

O modelo que melhor se ajustou é o da figura 6.8 com duas interacções estatisticamente significativas entre os dias de internamento e a idade depois de categorizados, e o sexo e a idade igualmente depois de categorizada.

Interessou igualmente avaliar o ajustamentos global do modelo, face ao modelo sem interacções.

Para isso recorremos ao comando *fitstat* já atrás descrito.

A saída computacional do mesmo encontra-se na figura 6.9 e permite concluir que este modelo é superior ao modelo inicial.

Note-se que a *odds* é de 1.88 no grupo com dias de internamento abaixo de 10, mas com o grupo etário mais elevado e no sexo a mudança para o sexo feminino, no mesmo grupo etário, apresenta uma odds de 0.44 portanto relevando menor multimorbilidade.

Foi este o modelo que utilizámos como modelo final

Da leitura desta figura ressalta as diferenças do **LR test**, do **AIC** e do **BIC** o que revela, como acima se disse, melhor ajustamento.

Repare-se igualmente na informação em rodapé, que confirma o que acabamos de dizer, e chama a atenção para a necessidade de nos assegurarmos do aninhamento dos modelos, o que é o caso.

```
. lintrend pmulti idade, group(10) plot(log)

The proportion and log odds of pmulti by categories of idade

  (Note: 10 idade categories of equal sample size;
     Uses mean idade value for each category)
```

idade	min	max	d	total	pmulti	logodds
2.1	0	11	506	23733	0.02	-3.83
21.5	11.25	27	463	23066	0.02	-3.89
31.6	27.25	35	669	25325	0.03	-3.61
39.6	35.2	44	1841	20972	0.09	-2.34
50.2	44.166667	55	6477	24966	0.26	-1.05
59.6	55.142857	63	9463	22969	0.41	-0.36
67.0	63.083333	70	11706	22957	0.51	0.04
73.5	70.125	76	13075	22470	0.58	0.33
79.4	76.125	82.66666	14183	22185	0.64	0.57
87.4	82.75	118	15791	23166	0.68	0.76

```
. lintrend pmulti dias_int, group(10) plot(log)

The proportion and log odds of pmulti by categories of dias_int

  (Note: 10 dias_int categories of equal sample size;
     Uses mean dias_int value for each category)
```

dias_int	min	max	d	total	pmulti	logodds
1	1	1	4105	23899	0.17	-1.57
1.979456	1.0454545	2	6146	41981	0.15	-1.76
2.962134	2.1111111	3	5288	33568	0.16	-1.68
3.938995	3.027027	4	5618	22457	0.25	-1.10
4.979199	4.03125	5.5	6549	18168	0.36	-0.57
6.468807	5.5333333	7	9813	23668	0.41	-0.34
8.775732	7.0322581	10	12282	23803	0.52	0.06
12.97253	10.090909	16	12375	22249	0.56	0.23
31.86512	16.2	364	11998	22016	0.54	0.18

Figura 6.6: Lintrend da idade e dias de internamento

```
. logistic pmulti sexo dsp adm_tip idade_gr dias_int_10

Logistic regression                                     Number of obs   =     231807
                                                        LR chi2(5)      =   53376.23
                                                        Prob > chi2     =     0.0000
Log likelihood = -118620.66                             Pseudo R2       =     0.1837

------------------------------------------------------------------------------
      pmulti | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        sexo |   .6018753   .0061715   -49.51   0.000     .5899001    .6140937
         dsp |   1.027354   .0012654    21.91   0.000     1.024877    1.029837
     adm_tip |   1.363825   .0153872    27.50   0.000     1.333998     1.39432
    idade_gr |   6.452289   .0678006   177.43   0.000     6.320761    6.586553
dias_int_10 |   2.170984    .025912    64.95   0.000     2.120787    2.222369
       _cons |   .2447621   .0059097   -58.29   0.000      .233449    .2566234
------------------------------------------------------------------------------
```

Figura 6.7: Modelo de efeitos principais

```
. logistic pmulti sexo dsp adm_tip idade_gr dias_int_10 dias_int_10#idade_gr sexo#idade_gr
note: 1.dias_int_10#0.idade_gr omitted because of collinearity
note: 1.dias_int_10#1.idade_gr omitted because of collinearity
note: 2.sexo#0.idade_gr omitted because of collinearity
note: 2.sexo#1.idade_gr omitted because of collinearity

Logistic regression                             Number of obs   =     231807
                                                LR chi2(7)      =   55676.94
                                                Prob > chi2     =     0.0000
Log likelihood = -117470.31                     Pseudo R2       =     0.1916
```

pmulti	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sexo	.442032	.005967	-60.48	0.000	.4304902	.4538832
dsp	1.025904	.0012453	21.07	0.000	1.023466	1.028347
adm_tip	1.371276	.0155579	27.83	0.000	1.34112	1.402111
idade_gr	5.956128	.1324052	80.27	0.000	5.702191	6.221374
dias_int_10	2.822183	.0447304	65.46	0.000	2.735861	2.911229
dias_int_10#idade_gr						
0 1	1.887198	.0433596	27.64	0.000	1.8041	1.974124
1 0	1	(omitted)				
1 1	1	(omitted)				
sexo#idade_gr						
1 1	.4486124	.0093216	-38.58	0.000	.4307094	.4672595
2 0	1	(omitted)				
2 1	1	(omitted)				
_cons	.3608789	.0099115	-37.11	0.000	.3419663	.3808375

Figura 6.8: O modelo com interacções

```
. fitstat, using(m0)

Measures of Fit for logistic of pmulti

                             Current             Saved       Difference
Model:                      logistic          logistic
N:                            231807            231807                0
Log-Lik Intercept Only    -145308.778       -145308.778            0.000
Log-Lik Full Model        -117470.309       -118620.661         1150.352
D                          234940.619(231799) 237241.322(231801)  2300.703(2)
LR                          55676.938(7)      53376.235(5)       2300.703(2)
Prob > LR                      0.000             0.000             0.000
McFadden's R2                  0.192             0.184             0.008
McFadden's Adj R2              0.192             0.184             0.008
ML (Cox-Snell) R2              0.214             0.206             0.008
Cragg-Uhler(Nagelkerke) R2     0.299             0.288             0.011
McKelvey & Zavoina's R2        0.281             0.265             0.016
Efron's R2                     0.237             0.228             0.008
Variance of y*                 4.576             4.474             0.102
Variance of error              3.290             3.290             0.000
Count R2                       0.760             0.767            -0.007
Adj Count R2                   0.249             0.271            -0.022
AIC                            1.014             1.023            -0.010
AIC*n                     234956.619        237253.322        -2296.703
BIC                       -2.629e+06        -2.626e+06        -2275.996
BIC'                      -55590.462        -53314.466        -2275.996
BIC used by Stata         235039.448        237315.444        -2275.996
AIC used by Stata         234956.619        237253.322        -2296.703

Difference of 2275.996 in BIC' provides very strong support for current model.

Note: p-value for difference in LR is only valid if models are nested.
```

Figura 6.9: Comparação dos modelos com e sem interacções

Capítulo 7

Discussão

7.1 Introdução

Com este trabalho pretendeu-se caracterizar a população com alta hospitalar na região de Lisboa e Vale do Tejo entre 2009 e 2011, com mais de 1 dia e menos de 365 dias de internamento, no que concerne á existência de padrões de diagnósticos que configurassem os conceitos de comorbilidade e/ou de multimorbilidade, conforme descritos na literatura amplamente citada no início do mesmo.

Para tal, recorreu-se aos dados dos GDH (Grupos de Diagnóstico Homogéneo) da ACSS, aos quais tivemos acesso.

Estes contém um conjunto de variáveis demográficas e códigos diagnósticos de acordo com a CID-9.

A partir destes códigos, em formato alfanumérico, construímos variáveis DUMMY (0/1) que representavam a presença, ou a ausência daquele diagnóstico.

Utilizámos então a ANÁLISE FACTORIAL, a partir de coeficientes de correlação tetratcóricos, para tentar resumir os dados e encontrar factores comuns que pudessem configurar grupos diagnósticos homogéneos.

Uma vez estes encontrados construímos uma variável que sintetiza a presença/ausência de multimorbilidade, considerada esta como a coexistência de 2 ou mais diagnósticos em qualquer dos factores isoladamente ou em simultâneo.

Essa variável foi então introduzida num modelo multivariado de REGRESSÃO LOGÍSTICA, como variável resposta, com um conjunto de covariáveis previamente seleccionadas.

Note-se que, em relação aos anos de 2009 e 2010 só tínhamos informação acerca dos episódios de internamento hospitalar e, não sobre os indivíduos

que tinham sido internados. A partir de 2011 foi introduzido no GDH uma variável que permite identificar os utentes por um numero fictício.

Tal facto permitiu poder agregar os episódios á volta dos indivíduos, embora não permita alocar objectivamente cada episódio a cada indivíduo num tempo determinado.

De qualquer forma este facto já permitiu fazer a agregação dos episódios e foi assim que procedemos á construção do modelo multivariado.

De realçar que retirámos da base de dados todos os diagnósticos que apresentavam uma frequência inferior a 1000 e igualmente todos os que se referiam aos códigos o eventos relacionados com a gravidez, o parto e o puerpério, não porque não possa existir multimorbilidade nessa área, mas porque ela é bastante específica e, portanto, para além do âmbito do nosso estudo.

Foram assim revistos 845928 episódios distribuídos da seguinte forma:

- 2009 - 307801
- 2010 - 306318
- 2011 - 231809

No ano de 2011 após a agregação dos episódios á volta do numero fictício de utente do GDH ficámos com 231809 observações.

Estudámos então o conjunto de variáveis *idade, sexo, dias de internamento, destino após alta, tipo de admissão* cujos resultados se apresentam pormenorizadamente no capítulo anterior.

Podemos resumidamente dizer que os resultados são bastante consistentes nos vários anos.

Resumimos os resultados para os três anos na figura 7.1

	2009		2010		2011	
	Med.	D.P.	Med.	D.P.	Med.	D.P.
Idade	48.43	27.9	48.64	28.29	50.82	6.12
Dias Int.	7.6	12.27	7.78	12.2	7.58	11.86
Sexol (\%)	Fem.	Mas.	Fem.	Mas.	Fem.	Mas.
	55.29	44.71	55.36	44.64	57.64	42.57
Dst. Ap. Alta (\%)	Dom.	Otr.	Dom.	Otr.	Dom.	Otr.
	92	8	92	8	92	8
Adm. Tipo (\%)	Urg.	Prg	Urg.	Prg	Urg.	Prg
	66.52	33.4	67.9	32.0	68.75	33.98

Figura 7.1: Resumo de variáveis

Verificamos que existem mais episódios em mulheres do que em homens, que os dias de internamento, embora não sendo muito longos com uma média a rondar os 7 dias, apresentam *outliers* marcados como aliás se pode constatar pela leitura dos gráficos de dispersão para cada ano apresentados no capítulo resultados nas figura 6.1, figura **??** e figura **??**.

Quanto ao destino após alta os doentes são maioritariamente encaminhados para o domicilio versus outros destinos.

Igualmente a admissão é feita através da urgência preferencialmente.

Considerando a análise bivariada, ao longo dos três anos, os resultados condizem igualmente bem.

De uma forma geral os homens são mais idosos, têm tempo de internamento mais prolongados, têm menos altas para o domicílio do que as mulheres, e têm mais óbitos.

Igualmente os episódios de urgência são mais frequentes nas mulheres.

Todos estes resultados se encontram discriminados no capítulo 7.

7.2 Análise Factorial

Passando ao campo da **Análise Factorial** verificamos que, de uma forma geral, os modelos apresentam baixos valores de explicação da variabilidade dos dados, sendo que com 8 factores se consegue explicar somente cerca de 30% da mesma.

Contudo devemos referir que este facto é usual em modelos com variáveis categóricas em que se verifica a presença de grande número de factores para atingir um bom nível de explicação.

Por outro lado, não nos podemos esquecer que estamos a lidar com episódios e não observações individuais, o que como já foi dito, provoca alguma distorção nos resultados.

Idealmente pretenderíamos poder realizar o estudo nas condições de um indivíduo um diagnóstico; contudo com os dados que dispomos tal não é possível, atendendo á forma como a base de dados está construída e também porque não é possível encontrar outra que satisfaça as nossas necessidades.

Apresentamos na figura 7.2 um quadro resumo dos vários factores definidos nos três anos, incluindo o de 2011 com os dados agregados.

Apesar de algumas diferenças entre os mesmos podemos constatar a presença de alguns factores de uma forma que poderíamos quase dizer constante.

Na realidade factores como o que chamaremos cardiometabólico (incluindo diabetes, dislipidemia, obesidade, HTA, EAM), o da Doença cere-

2009	2010	2011	2011agr
Neo da prostata; p. neurologicas e demenciais; d. hipertensiva;d. resp.; d. renais e urinarios	Abuso alc. drog. d. figado d. musc. esquel.	Diabetes dislip. obesid. Parkinson demencias	D. alcool d. hepatica
Idem; compl. actos med e cir.	D. Respir. inc. neo pulm	D. isquem cor. D. respir.	D. renal D. Cardiorenal
Hta,;diabetes;obesidade;EAM	D. neuro (Parkinson, demencia D. Cere. vasc)	D. gastrointestinal	Hta diabetes EAM ICC
D. figado inc. cirrose	EAM D. coronar D. cerebro v. Insuf. Renal Ag.	D. osteoartic.	D. cerebrovasc.
D. gen. fem exc. grav e parto	HTA diabetes displip. d. corn. ICC	Ac. proc. clinic. sepsis, d. vias urinar.	D. osteoarticul.
Neoplasias	D. Digest. exc. figado	Nega. fractur.	D. respirator. inc. pneumonia
Dep. alcool drogas e D. Inf. Ap. Respir.	Neoplasias	Neoplasias	D. gerais neo seund.
D. Cerebrovasc e Depress.	D. cerebrovasc. demencias	D. neuromusc.	

Figura 7.2: Resumo dos factores

brovascular, o das Doenças hepáticas e o osteomuscular encontram-se presente nos vários anos, pese embora o facto dos *loadings* serem diferentes nos vários anos.

Para além das flutuações populacionais, o facto de estarmos a lidar com episódios terá eventualmente importância neste facto.

Recordemos que os modelos com episódios se devem interpretar de uma forma mais indicativa que definitiva, uma vez que o método de **Análise Factorial** se destina a observações individuais.

Quando passamos ao nível individual as coisas apresentam-se mais claras.

Assim o factor 1 ligado ao consumo de álcool e ás doenças do fígado tem um *loading* de 5.13 explicando cerca de 33% da variabilidade.

Os factores 2 e 3 apresentam *loadings* mais baixos (2.07 e 1.41) explicando em conjunto cerca de 23% de variabilidade.

É para nós um pouco estranho a razão pela qual o factor cardiometabólico tem um valor tão baixo. Mas cremos que a razão deve estar no tipo de dados.

Concluímos assim que, embora os nossos resultados se não sejam muito diferentes dos da literatura , existe um posicionamento diferente devido, sem dúvida á natureza dos dados.

Esta constitui a diferença fundamental entre o nosso trabalho e o dos outros estudos citados.

Assim no estudo de Cornell [14] foram analisados doentes de cuidados de saúde primários (N=1327328); no estudo de Miera [23] foram incluídos 4310 indivíduos com alta hospitalar. O estudo de Schaffer englobou 149280 indivíduos recolhidos de uma base de dados de uma companhia alemã [63]. Finalmente no estudo de Newcomer [54] foram incluidos 15480 doentes provenientes de uma organização de cuidados de saúde.

Em todos estes estudos e, muitos mais se poderiam referir, foram sempre utilizados casos individuais.

De acordo com uma revisão recente de Haregu em 2012 [30] foram passados em revista os aspectos principais das perspectivas, constructos e métodos de medir a multimorbilidade.

A metodologia de análise destes estudos centrou-se quer na Análise de Clusters, quer na **Análise Factorial**.

Recentemente foram publicados mais trabalhos, nomeadamente de Prados-Torres et all. [56] e de Kircheberger [42] em que a Análise Factorial foi utilizada para medir a multimorbilidade, embora em diferentes populações.

Em todos estes estudos se verificou a presença dos vários grupos resumidos na figura 7.3

Note-se que, embora com variações os grupos que se encontram nos vários estudos são relativamente semelhantes, pese embora a base populacional dos mesmos (hospitais, cuidados primários, etc.) e a diferente técnica usada.

Outro aspecto que se deve salientar é que, todos estes estudos partiram de uma lista de diagnósticos previamente estabelecida, quer a partir de listas de consenso, painéis de peritos, etc.

No nosso caso não fizemos nenhuma selecção prévia, ficando assim mais expostos ás variações da composição da própria base de dados.

Contudo a presença das varias patologias parece bastante consistente, com particular ênfase para o grupo metabólico-cardiovascular, mas também

Estudo de Cornell (Clusters)	Estudo de Miera (Clusters)	Estudo de Schaffer (Factorial)	Estudo de Newcomer (Clusters)
Obesidade	Descritivo sem construção de grupos	Cardiovasc. com doença metabolica	Cardiometabolico
Metabólico		Cardiovasc. com doença cerebrovascular	Doença renal e diabetes
Neurovascular		Cluster ansiedade/ depressão	Cirurgia ortopédica/ abdominal com obesidade
Fígado			Doença mental e obesidade
Diagnóstico Dual			
Misto			

Figura 7.3: Resumo dos estudos citados

para o grupo da doença cerebrovascular e até para as neoplasias que não parecem tão evidentes nos outros estudos.

7.3 Regressão logística

O nosso modelo de regressão logística foi apresentado no capítulo anterior.

Os resultados obtidos não parecem indicar que este modelo se ajuste particularmente bem aos dados.

Dos vários modelos testados, e cujo resultado não apresentamos, ficamos com aquele que cumpria as condições acima definidas (parcimónia e sentido biológico) conforme descrito; este foi estudado "contra" o modelo original (sem interacções) a que chamámos *m0* .

O resultado do comando *fitstat* apresentado na figura 6.9 indica um a diferença estatisticamente significativa entre os dois modelos, favorável ao modelo final (**AIC** mais baixo).

Observamos que todos os coeficientes se apresentam estatisticamente diferentes de 0, pelo terão de ser considerados no modelo.

Nota-se igualmente o impacto do grupo etário com uma odds de 6.45, bem como dos dias de internamento mais prolongado com uma odds de

2.82.

Como já tínhamos suspeitado da análise das tabelas do *lintrend* estas duas variáveis são de importância capital.

O tipo de admissão apresenta igualmente importância com uma odds de 1.37 aumentando assim com a passagem da mesma de programada a urgente.

Mantendo as outras variáveis inalteradas existe uma interacção entre o sexo masculino e os dias de internamento, bem como entre os sexo masculino e o grupo etário mais avançado, ambas estatisticamente significativas *(pvalue>0.00)* (ver figura 6.8).

Igualmente a introdução da interacção modifica a odds do grupo etário em cerca de 8%, o que sendo um valor relativamente baixo, denota algum possível confundimento.

Note-se que noutros modelos testados esta variação é bem maior, o que leva a crer que a interacção entre os dias de internamento e o grupo etário tem algum efeito de confundimento na relação entre a multimorbilidade e o próprio grupo etário.

Com uma variação no **AIC** de -2188.635 e no **BIC** de -2167.925, ambas estatisticamente significativas o modelo é o mais ajustado dos testados.

Note-se igualmente a diferença entre a desvio com um **LR-test** estatisticamente significativo com um *p-value>0.00* correspondente.

Uma nota final para a necessidade dos modelos se encontrarem aninhados (o que é o caso), condição para que estas estatísticas sejam válidas.

Contudo devemos, mais uma vez, enfatizar a natureza dos nossos dados não nos permitirá ir muito longe.

A modelação estatística parece indicar que mantendo os outros valores constantes a idade mais avançada (OR:5.95), e a duração do internamento superior a 10 dias (OR:2.82) têm manifesta relação com a ocorrência da multimorbilidade, como descrita (2 ou mais diagnósticos em qualquer dos factores encontrados).

Igualmente a interacção entre os dias de internamento e o grupo etário mais avançado bem como o sexo (masculino) e o grupo etário mais avançado contribuem igualmente para a ocorrência da multimorbilidaqde.

Não se torna possível chegar, com rigor, muito mais longe na análise destes resultados.

Como se disse tal facto advém da natureza dos dados utilizados.

7.3.1 Aplicação de técnicas de diagnóstico

Conforme explicitámos no capítulo 5 fomos agora proceder á avaliação do correcto ajustamento do modelo á eficácia da caracterização da variável

resposta *pmulti* face ao conjunto das covariáveis seleccionadas.

Começámos por utilizar o teste de Hosmer and Lemeshow, conforme descrito, e depois passámos á análise gráfica das medidas propostas pelos autores (Hosmer and Lemeshow, op. citada).

O resultado do teste está presente na figura 7.4

```
Logistic model for pmulti, goodness-of-fit test

(Table collapsed on quantiles of estimated probabilities)
(There are only 9 distinct quantiles because of ties)

      number of observations =    231807
            number of groups =         9
     Hosmer-Lemeshow chi2(7) =      4319.56
                 Prob > chi2 =         0.0000
```

Figura 7.4: O teste de Hosmer e Lemshow

A leitura deste quadro permite-nos verificar que o nosso modelo não passa no respectivo teste.

Observamos igualmente que o mesmo nos indica a existência de empates que não permitem construir os 10 grupos propostos pelos autores para realização do teste.

A razão deste facto prende-se seguramente com a natureza dos dados e da forma como é feito o processo de agregação, como atrás foi dito.

Na figura 7.5 apresentamos os gráficos de $\widehat{\pi} vs \chi_i^2$, $\widehat{\pi} vs \Delta_i$ e $\widehat{\pi} vs \widehat{\beta}_i$ correspondentes aos gráficos analíticos propostos por Hosmer e Lemeshow e atrás citados.

Como vemos estes gráficos não se apresentam com o aspecto tradicional das curvas que se cruzam exemplificando assim a variação homogénea da probabilidade; pelo contrário esta varia pouco e existem outliers seguramente influentes no modelo.

Parafraseando Hosmer and Lemeshow (op. cit. 178–179), "os pontos vindo do canto superior esquerdo para o canto inferior direito correspondem a padrões de covariatos com o número de resultados positivos igual ao número no grupo; os pontos na outra curva correspondem a 0 resultados positivos".

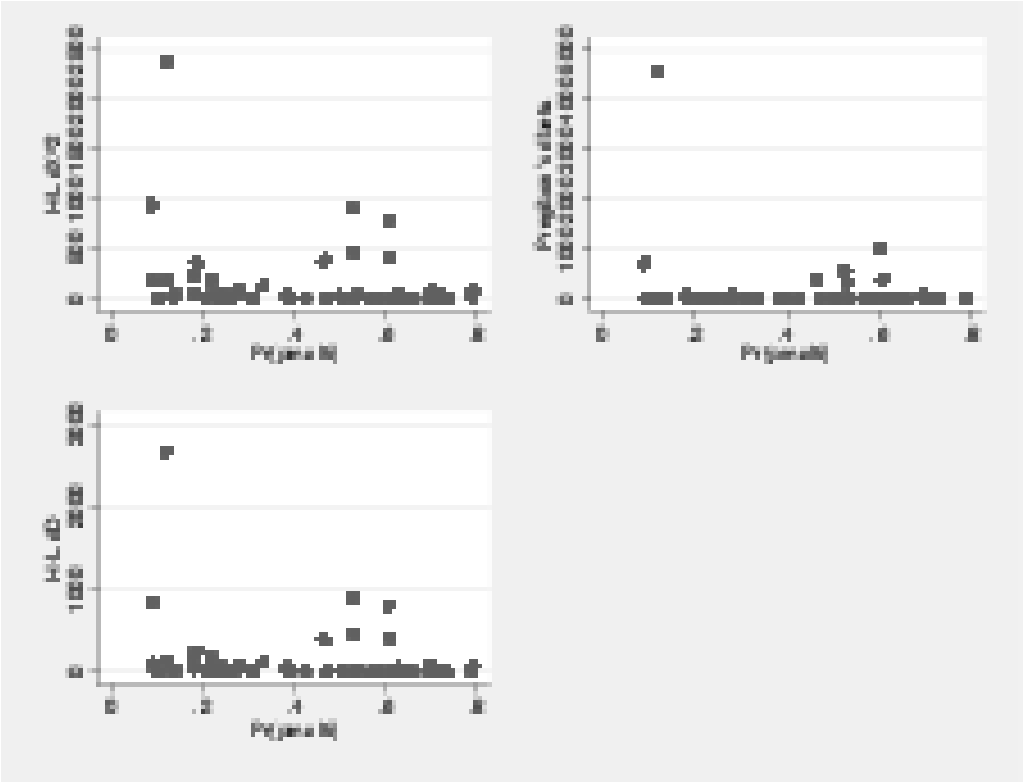


Figura 7.5: Os gráficos de diagnóstico

Sendo a maioria dos padrões de covariatos únicos os pontos tendem a dispor-se ao longo de uma ou outra das curvas; os pontos fora das curvas correspondem a padrões de covariatos repetidos.

Ora esse é precisamente o nosso caso; supomos que a agregação dos episódios, feita da forma utilizada, conduza a um elevado número de empates.

Mais uma vez afirmamos que estamos convictos que a natureza dos dados seja a responsável por esta situação o que nos levará a desejar prosseguir este trabalho noutras condições

Capítulo 8

Conclusões

O estudo que levámos a efeito recorrendo ao método de Análise Factorial sobre uma base de dados do GDH de doentes com alta hospitalar (considerando nestas os falecidos) entre 2009 e 2011 com mais de 1 dia de internamento e menos de 365 dias excluindo os da ala obstétrica, pretendia encontrar padrões de multimorbilidade nesses internamentos.

Inicialmente pensou-se em termos de comorbilidade da forma como foi inicialmente definida por [22].

Contudo a evolução do próprio trabalho levou-nos a perceber bem cedo que o conceito deveria ser mais alargado na linha do pensamento posterior de outros autores como van den Akker [69].

Na realidade fazia sentido, para nós, encarar os episódios de doença não a partir de uma doença índice, mas sim tentando abarcar toda a realidade das polipatologias que constituem a base da multimorbilidade.

Assim seguindo o curso da análise viemos a encontrar valores de multimorbilidade que se situaram entre os 31 e os 37% nos três anos.

O impacto deste conhecimento é grande considerando duas vertentes já anteriormente referidas:

- Por uma lado medir a prevalência considerando as doenças de uma forma isolada, ou no seu conjunto de multimorbilidade é particularmente diferente. Uma coisa será a diabetes, a hipertensão, a obesidade de per si, por exemplo, outra considerar as três doenças integrando um padrão de multimorbilidade

- Por outro lado planear intervenções terapêuticas, ou estabelecer "guidelines" para o tratamento das mesmas apresenta cenários diferentes se considerarmos o doente portador de cada patologia individualmente, ou o doente portador de várias patologias, como bem lembra Daniel Campbell-Scherer [9]

No ano de 2011, a partir de uma alteração do GDH, tivemos a possibilidade de associar ao mesmo paciente um ou vários episódios, muito embora não tivéssemos possibilidade de avaliar a temporalidade dessa associação, a sua ordem cronológica, ou mesmo o estabelecimento em que se efectivou o internamento.

Isto constitui uma dificuldade para uma identificação precisa dos diversos episódios, e da sua relação com o os dentes que os sofrem, gerando igualmente várias situações de empate nas observações, o que fragiliza os dados e prejudica a análise. Contudo existe semelhança entre os nossos valores e os encontrados na literatura, quer quantitativa como qualitativamente; isto leva-nos a pensar que, mesmo assim, talvez não estejamos demasiado longe da verdade.

Os grupos diagnósticos associando doenças metabólicas como a diabetes a obesidade e a dislipidemia com a hipertensão e a doença coronária aguda, o grupo hepático, o grupo reno vascular e o cerebrovascular, bem como o grupo das doenças do sistema nervoso são fortemente compatíveis com aquilo que outros investigadores encontraram usando o mesmo método ou outros.

Existem contudo, também, algumas diferenças.

No nosso trabalho o peso do *cluster psiquiátrico* (incluindo psicoses, demência, ansiedade, etc.) apresentou-se menos importante do que noutros estudos, nomeadamente no de Schaffer [63] entre outros [15].

Por outro lado o grupo das neoplasias teve sempre mais expressão no nosso trabalho; as razões para tal não podem ser aqui cabalmente apuradas, pelo que teremos de esperar por estudos posteriores para esclarecer melhor estas diferenças.

Quanto ao modelo estatístico de regressão logística, embora os resultados pareçam bastante concordantes com o conhecimento prático, idade mais avançada, mais dias de internamento, sexo masculino, mais multimorbilidade, não foi possível demonstrá-lo cabalmente, provavelmente devido á natureza dos dados utilizados.

A utilização do GDH apresentou portanto alguns desafios.

A nível da codificação das doenças esta não será uniforme a nível de todos os codificadores, o que introduzirá diferenças que se reflectirão na homogeneidade final dos diagnósticos. Lembremos que os códigos CID-9 são de forma caule e folhas, existindo para um diagnóstico base um conjunto de sub-diagnósticos relacionados.

Assim e uma vez que, no fundo, são utilizados para pagamento dos serviços prestados é possível que se tente apresentar o máximo de códigos que possam corresponder a uma melhor compensação.

Desta forma os GDH, sendo neste momento a ferramenta possível de utilizar, apresenta limitações que gostaríamos de ultrapassar no futuro. Podemos assim concluir que:

- Encontrámos multimorbilidade entre 31 e 37% nos doentes com alta hospitalar na Região de Lisboa e Vale do Tejo
- Os dados atualmente disponíveis pelo menos a partir da fonte usada, apresentam alguma limitações que dificultam uma análise aprofundada e qualificada do problema
- De qualquer forma a comparação com os resultados da literatura, pese a diferença de fontes de dados utilizadas, são relativamente consistentes
- Supomos que este trabalho pode ter continuidade se se disponibilizarem condições para se construir uma base de dados que possa recolher informação com as características necessárias á realização de um estudo mais aprofundado

Acreditamos que os métodos utilizados conferiram resultados quantitativos e qualitativos de boa qualidade, tendo ficado demonstrado, de uma forma positiva, serem uma ferramenta viável e consistente para enfrentar aos desafios do estudo da co- e da mutilmorbilidade.

Bibliografia

[1] ACSS. Grupos de diagnostico homogeneo. "http://portalcodgdh.min.saude.pt".

[2] Alan Agresti. *Categorical data analysis*, volume 359. John Wiley & Sons, 2002.

[3] M Aragona. The role of comorbidity in the crisis of the current psychiatric classification system. *Philosophy, Psychiatry, & Psychology*, 16(1):1–11, 2009.

[4] Wilbert S Aronow, Chul Ahn, Anthony D Mercando, Stanley Epstein, et al. Prevalence of coronary artery disease, complex ventricular arrhythmias, and silent myocardial ischemia and incidence of new coronary events in older persons with chronic renal insufficiency and with normal renal function. *The American journal of cardiology*, 86(10):1142, 2000.

[5] Joseph Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3):47–53, 1946.

[6] T. Jean Blocker and Douglas Lee Eckbert. Environemental issues and womens issues; general concerns and local hazards. *Social Sciences Quarterly*, 70(3):586–593, 1989.

[7] Cynthia M Boyd, Martin Fortin, et al. Future of multimorbidity research: How should understanding of multimorbidity inform health system design. *Public Health Reviews*, 32(2):451–474, 2010.

[8] Sharon G Bruce, Natalie D Riediger, James M Zacharias, and T Kue Young. Peer reviewed: Obesity and obesity-related comorbidities in a canadian first nation population. *Preventing chronic disease*, 8(1), 2011.

[9] Denise Campbell-Scherer. Multimorbidity: a challenge for evidence-based medicine. *Evidence Based Medicine*, 15(6):165–166, 2010.

[10] Raymond B Cattell. The scree test for the number of factors. *Multivariate behavioral research*, 1(2):245–276, 1966.

[11] Gillian E Caughey, Emmae N Ramsay, Agnes I Vitry, Andrew L Gilbert, Mary A Luszcz, Philip Ryan, and Elizabeth E Roughead. Comorbid chronic diseases, discordant impact on mortality in older people: a 14-year longitudinal population study. *Journal of epidemiology and community health*, 64(12):1036–1042, 2010.

[12] Bernard R Chaitman, Lloyd D Fisher, Martial G Bourassa, Kathryn Davis, William J Rogers, Charles Maynard, Denis H Tyras, Robert L Berger, Melvin P Judkins, Ivar Ringqvist, et al. Effect of coronary bypass surgery on survival patterns in subsets of patients with left main coronary artery disease: report of the collaborative study in coronary artery surgery (cass). *The American journal of cardiology*, 48(4):765–777, 1981.

[13] Mary E Charlson, Peter Pompei, Kathy L Ales, and C Ronald MacKenzie. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of chronic diseases*, 40(5):373–383, 1987.

[14] John E Cornell, Jacqueline A Pugh, John W Williams Jr, Lewis Kazis, Austin FS Lee, Michael L Parchman, John Zeber, Thomas Pederson, Kelly A Montgomery, Polly Hitchcock Noël, et al. Multimorbidity clusters: Clustering binary data from multimorbidity clusters: Clustering binary data from a large administrative medical database. *Applied Multivariate Research*, 12(3):163–182, 2009.

[15] Vincent de Groot, Heleen Beckerman, Gustaaf J Lankhorst, and Lex M Bouter. How to measure comorbidity: a critical review of available methods. *Journal of clinical epidemiology*, 56(3):221–229, 2003.

[16] Prakash C Deedwania. Mechanism of the deadly quartet. *The Canadian journal of cardiology*, 16:17E, 2000.

[17] Miguel Delgado-Rodríguez and Javier Llorca. Bias. *Journal of Epidemiology and Community Health*, 58(8):635–641, 2004.

[18] Richard A Deyo, Daniel C Cherkin, and Marcia A Ciol. Adapting a clinical comorbidity index for use with icd-9-cm administrative databases. *Journal of clinical epidemiology*, 45(6):613–619, 1992.

[19] William D'Hoore, André Bouckaert, and Charles Tilquin. Practical considerations on the use of the charlson comorbidity index with administrative data bases. *Journal of clinical epidemiology*, 49(12):1429–1433, 1996.

[20] Anne Elixhauser, Claudia Steiner, D Robert Harris, and Rosanna M Coffey. Comorbidity measures for use with administrative data. *Medical care*, 36(1):8–27, 1998.

[21] REIS Elizabete. Estatística multivariada aplicada, 1997.

[22] Alvan R. Feinstein. The pre-therapeutic classification of co-morbidity in chronic disease. *Journal of Chronic Diseases*, 23(7):455–468, 1970.

[23] Manuel Francisco Fernández Miera. Patients with multimorbidity in the hospital setting. *Gaceta Sanitaria*, 22(2):137–141, 2008.

[24] Martin Fortin, Lise Lapointe, Catherine Hudon, and Alain Vanasse. Multimorbidity is common to family practice: is it commonly researched? *Canadian Family Physician*, 51(2):244–245, 2005.

[25] Martin Fortin, Lise Lapointe, Catherine Hudon, Alain Vanasse, Antoine L Ntetu, and Danielle Maltais. Multimorbidity and quality of life in primary care: a systematic review. *Health and Quality of life Outcomes*, 2(1):51, 2004.

[26] Richard L. Gorsuch. *Factor Analysis*. L. Erlbaum Associates, 1983.

[27] Sheldon Greenfield, Giovanni Apolone, Barbara J McNeil, Paul D Cleary, et al. The importance of co-existent disease in the occurrence of postoperative complications and one-year recovery in patients undergoing total hip replacement. comorbidity and outcomes after hip replacement. *Medical care*, 31(2):141, 1993.

[28] Dianne L Groll, Teresa To, Claire Bombardier, and James G Wright. The development of a comorbidity index with physical function as the outcome. *Journal of clinical epidemiology*, 58(6):595–602, 2005.

[29] Lawrence C Hamilton. *Regression with graphics: A second course in applied statistics*, volume 1. Duxbury Press Belmont, CA, 1992.

[30] TN Haregu, B Oldenburg, G Setswe, and J Elliott. Perspectives, constructs and methods in the measurement of multimorbidity and comorbidity: A critical review. *The Internet Journal of Epidemiology*, 10(2), 2012.

[31] Charles Hennekens and Julie Buring. *Epidemiology in medicine*, volume 515. Lippincott Williams & Wilkins, 1987.

[32] Miguel A Hernan, Sonia Hernandez-Diaz, and James M Robins. A structural approach to selection bias. *Epidemiology*, 15(5):615–625, 2004.

[33] Oliver Hirsch, Stefan Bösner, Eyke Hüllermeier, Robin Senge, Krzysztof Dembczynski, and Norbert Donner-Banzhoff. Multivariate modeling to identify patterns in clinical data: the example of chest pain. *BMC medical research methodology*, 11(1):155, 2011.

[34] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. Wiley. com, 2013.

[35] C Hudon, M Fortin, and A Vanasse. Cumulative illness rating scale was a reliable and valid index in a family practice context. *Journal of clinical epidemiology*, 58(6):603–608, 2005.

[36] James I Hudson, Don L Goldenberg, Harrison G Pope, Paul E Keck, and Lynn Schlesinger. Comorbidity of fibromyalgia with medical and psychiatric disorders. *The American journal of medicine*, 92(4):363–367, 1992.

[37] Kyoko Imamura, Morag McKinnon, Robert Middleton, and Nick Black. Reliability of a comorbidity measure: the index of co-existent disease (iced). *Journal of clinical epidemiology*, 50(9):1011, 1997.

[38] Robert John, Dave S Kerby, and Catherine Hagan Hennessy. Patterns and impact of comorbidity and multimorbidity among community-resident american indian elders. *The Gerontologist*, 43(5):649–660, 2003.

[39] Moreson H Kaplan and Alvan R Feinstein. The importance of classifying initial co-morbidity in evaluating the outcome of diabetes mellitus. *Journal of chronic diseases*, 27(7):387–404, 1974.

[40] Jae On Kim and Charles W Mueller. Introduction to factor analysis: What it is and how to do it (quantitative applications in the social sciences# 13). beverly hills, 1982.

[41] Jae-on Kim and Charles W Mueller. *Statistical methods and practical issues*. Sage Publications, 1982.

[42] Inge Kirchberger, Christa Meisinger, Margit Heier, Anja-Kerstin Zimmermann, Barbara Thorand, Christine S Autenrieth, Annette Peters, Karl-Heinz Ladwig, and Angela Döring. Patterns of multimorbidity in the aged population. results from the kora-age study. *PloS one*, 7(1):e30556, 2012.

[43] Stanislav Kolenikov and Gustavo Angeles. The use of discrete data in principal component analysis for socio-economic status evaluation, 2005.

[44] Karen Kuhlthau, Timothy GG Ferris, Anne C Beal, Steven L Gortmaker, and James M Perrin. Who cares for medicaid-enrolled children with chronic conditions? *Pediatrics*, 108(4):906–912, 2001.

[45] BS Linn, MW Linn, L Gurel, et al. Cumulative illness rating scale. *Journal of the American Geriatrics Society*, 16(5):622, 1968.

[46] Mark S Litwin, Sheldon Greenfield, Eric P Elkin, Deborah P Lubeck, Jeanette M Broering, and Sherrie H Kaplan. Assessment of prognosis with the total illness burden index for prostate cancer. *Cancer*, 109(9):1777–1783, 2007.

[47] J Scott Long. *Regression models for categorical and limited dependent variables*, volume 7. Sage, 1997.

[48] J Scott Long and Jeremy Freese. Regression models for categorical dependent variables using stata. *Stata Press books*, 2006.

[49] Brian MacMahon, Thomas F Pugh, et al. *Epidemiology: principles and methods.* Boston: Little Brown & Co. Published in Great Britain by J. & A. Churchill, London., 1970.

[50] Alessandra Marengoni, Debora Rizzuto, Hui-Xin Wang, Bengt Winblad, and Laura Fratiglioni. Patterns of chronic multimorbidity in the elderly population. *Journal of the American Geriatrics Society*, 57(2):225–230, 2009.

[51] João Maroco. *Análise estatística: com utilização do SPSS*. 2003.

[52] P Allison Minugh, Ted D Nirenberg, Patrick R Clifford, Richard Longabaugh, Bruce M Becker, and Robert Woolard. Analysis of alcohol use clusters among subcritically injured emergency department patients. *Academic emergency medicine*, 4(11):1059–1067, 1997.

[53] Dana C Miskulin, Nicolaos V Athienites, Guofen Yan, Alice A Martin, Daniel B Ornt, John W Kusek, Klemens B Meyer, and Andrew S Levey. Comorbidity assessment using the index of coexistent diseases in a multicenter clinical trial. *Kidney international*, 60(4):1498–1510, 2001.

[54] Sophia R Newcomer, John F Steiner, Elizabeth A Bayliss, et al. Identifying subgroups of complex patients with cluster analysis. *The American journal of managed care*, 17(8):e324, 2011.

[55] T Pincus, LF Callahan, et al. Taking mortality in rheumatoid arthritis seriously–predictive markers, socioeconomic status and comorbidity. *The Journal of rheumatology*, 13(5):841, 1986.

[56] Alexandra Prados-Torres, Beatriz Poblador-Plou, Amaia Calderón-Larrañaga, Luis Andrés Gimeno-Feliu, Francisca González-Rubio, Antonio Poncel-Falcó, Antoni Sicras-Mainar, and José Tomás Alcalá-Nalvaiz. Multimorbidity patterns in primary care: interactions among chronic diseases using factor analysis. *PloS one*, 7(2):e32190, 2012.

[57] Patrick S Romano, Leslie L Roost, and James G Jollis. Presentation adapting a clinical comorbidity index for use with icd-9-cm administrative data: differing perspectives. *Journal of clinical epidemiology*, 46(10):1075–1079, 1993.

[58] Kenneth J Rothman, Sander Greenland, and Timothy L Lash. *Modern epidemiology*. Lippincott Williams & Wilkins, 2008.

[59] Peter J Rousseeuw and Leonard Kaufman. Finding groups in data: An introduction to cluster analysis. *John, John Wiley & Sons*, 1990.

[60] Renzo Rozzini, Giovanni B Frisoni, Luigi Ferrucci, Piera Barbisoni, Toni Sabatini, Piera Ranieri, Jack M Guralnik, and Marco Trabucchi. Geriatric index of comorbidity: validation and comparison with other measures of comorbidity. *Age and ageing*, 31(4):277–285, 2002.

[61] Jane S Saczynski, Jerry H Gurwitz, Sandhyasree Padmanabhan, Robert J Goldberg, David J Magid, David H Smith, Sue Hee Sung, and Alan S Go. Patterns of complex comorbidity in older patients with heart failure. 2011.

[62] William C Sanderson, Aaron T Beck, and Judith Beck. Syndrome comorbidity in patients with major depression. *Am J Psychiatry*, 147:1025–1028, 1990.

[63] Ingmar Schafer, Eike-Christin von Leitner, Gerhard Schön, Daniela Koller, Heike Hansen, Tina Kolonko, Hanna Kaduszkiewicz, Karl Wegscheider, Gerd Glaeske, and Hendrik van den Bussche. Multimorbidity patterns in the elderly: a new approach of disease clustering identifies complex interrelations between chronic conditions. *PLoS One*, 5(12):e15941, 2010.

[64] Florian Schneider, Vladimir Kaplan, Roksana Rodak, Edouard Battegay, and Barbara Holzer. Prevalence of multimorbidity in medical inpatients. *Swiss medical weekly*, 142:w13533, 2012.

[65] Barbara Starfield, Klaus W Lemke, Terence Bernhardt, Steven S Foldes, Christopher B Forrest, and Jonathan P Weiner. Comorbidity: implications for the importance of primary care in 'case'management. *The Annals of Family Medicine*, 1(1):8–14, 2003.

[66] John S. Uebersax. Iintroduction to the tetrachoric and polychoric ccorrelation coeficients.

[67] Jose M Valderas, Barbara Starfield, Bonnie Sibbald, Chris Salisbury, and Martin Roland. Defining comorbidity: implications for understanding health and health services. *The Annals of Family Medicine*, 7(4):357–363, 2009.

[68] Gerald Van Belle, Lloyd D Fisher, Patrick J Heagerty, and Thomas Lumley. *Biostatistics: a methodology for the health sciences*, volume 519. Wiley. com, 2004.

[69] Marjan van den Akker, Frank Buntinx, and J André Knottnerus. Comorbidity or multimorbidity. *European Journal of General Practice*, 2(2):65–70, 1996.

[70] Hendrik van den Bussche, Daniela Koller, Tina Kolonko, Heike Hansen, Karl Wegscheider, Gerd Glaeske, Eike-Christin von Leitner, Ingmar Schäfer, and Gerhard Schön. Which chronic diseases and disease combinations are specific to multimorbidity in the elderly? results of a claims data based cross-sectional study in germany. *BMC public health*, 11(1):101, 2011.

[71] Peter T Weir, Gregory A Harlan, Flo L Nkoy, Spencer S Jones, Kurt T Hegmann, Lisa H Gren, and Joseph L Lyon. The incidence of fibromyalgia and its associated comorbidities: a population-based retrospective cohort study based on international classification of diseases, 9th revision codes. *JCR: Journal of Clinical Rheumatology*, 12(3):124, 2006.